56

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Facultad de Ciencias

El estudio de la deserción en una Universidad Pública en México: El caso de la Universidad Autónoma Metropolitana, Unidad Iztapalapa, División de Ciencias Básicas e Ingeniería

T E S I S

Que para obtener el título de

ACTUARIA

presenta

CLAUDIA ELI LOZADA CAN

FACULTAD DE CIENCIAS
UNAM

Director de tesis: M. en C. Inocencio Rafael Madrid Ríos

DIVISION DE ESTUDIOS PROFESIONALES
2001
FACULTAD DE CIENCIAS
SECCION ESCOLAR

2001

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

DERECHOS RESERVADOS © PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

VNIVERSIDAD NACIONAL
AVPNOMA DE
MEXICO

M. en C. ELENA DE OTEYZA DE OTEYZA
Jefa de la División de Estudios Profesionales
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:

"El estudio de la deserción en una Universidad Pública en México: el caso de la Universidad Autónoma Metropolitana, Unidad Iztapalapa, División de Ciencias Básicas e Ingeniería"

realizado por: Claudia Elí Lozada Can.

Con número de cuenta 9650367-4, pasante de la carrera de Actuaría.

Dicho trabajo cuenta con nuestro voto aprobatorio.

A t e n t a m e n t e

Director de tesis Propietario	M. en C. Inocencio Rafael Madrid Ríos
Propietario	Mat. Margarita Elvira Chávez Cano
Propietario	Dr. Ignacio Méndez Ramírez
Suplente	M. en A. P. María del Pilar Alonso Reyes
Suplente	M. en C. Francisco Sánchez Villarreal

Consejo Departamental de Matemáticas.

P.A.
M. en C. José Antonio Flores Díaz.

A mi papá Manuel Lozada por su tenaz apoyo, orientación, ejemplo y cariño para procurar mi superación. Por ser el eje central de mi existencia y ser responsable de cada uno de los peldaños alcanzados.

A mi mamá Socorrito por el afecto, entereza, apoyo y amor de madre. Por su valentía y apoyo incondicional en cada momento. Te quiero mucho.

A mi hermano Mario por su solidaria complacencia, sensibilidad y protección.

A la Maestra Magdalena Fresán quien, además de amiga, es la impulsora de mi interés por realizar la tesis en este tema. Además, me proporcionó las herramientas necesarias para establecer la estructura de la investigación y en la obtención de los insumos que le dan forma. Fue un elemento vital en el análisis y elaboración de los planteamientos. Sin su generosa asesoría, no habría sido posible la realización de este trabajo.

Muchas Gracias

Al M. en C. Inocencio Rafael Madrid Ríos por su participación, colaboración y apoyo en este proyecto.

Al Dr. Ignacio Méndez Ramírez por sus valiosos comentarios que hicieron posible la culminación de este trabajo.

A los sinodales: Mat. Margarita Chávez Cano, M. en AP. Pilar Alonso Reyes y M. en C. Francisco Sánchez Villareal por su colaboración en la realización de este proyecto.

A la Maestra Alejandra Romo por su generosa dedicación en la revisión del texto y por su amistad.

A Mary Carmen Silva por su gran respaldo logístico, protección y participación en mí proyecto de vida.

A Paty Acuña por su interés en mí persona.

A Lulú y Carlos por su constructiva compañía en los momentos apremiantes y por su cálida amistad.

A ti por ser tan especial y estar junto a mí.

A todas las personas que me brindaron su apoyo, confianza y amistad durante la realización de este trabajo.

El estudio de la deserción en una Universidad Pública en México: el caso de la Universidad Autónoma Metropolitana, Unidad Iztapalapa, División de Ciencias Básicas e Ingeniería.

CAPÍTULO 3. ANÁLISIS DE DATOS

CAPÍTULO 4. DISCUSIÓN

CAPITULO 5. CONCLUSIONES

Glosario

Anexos

El estudio de la deserción en una Universidad Pública en México: el caso de la Universidad Autónoma Metropolitana, Unidad Iztapalapa, División Ciencias Básicas e Ingeniería

Introducción

La idea de este trabajo surge de la necesidad de realizar estudios de deserción estudiantil en el sistema de educación superior en México. El objetivo general de esta investigación consiste en detectar oportunamente estudiantes con alto riesgo de abandonar sus estudios. La metodología utilizada *permitió construir un modelo de análisis y diagnóstico de la trayectoria escolar de los* estudiantes, a través de las diversas etapas que comprenden los planes de estudio de licenciatura en las instituciones de educación superior públicas. El modelo construido permite estimar la probabilidad de que un estudiante no abandone la universidad durante los diferentes periodos lectivos que comprenden los planes de estudio.

La hipótesis que orienta este trabajo es la siguiente: El género del estudiante, la edad al *ingreso, algunos factores relacionados con los antecedentes educativos* y los resultados del examen de admisión, influyen en la permanencia de los estudiantes en la universidad. A partir de ellos es posible determinar el riesgo que tiene un estudiante de abandonar o de rezagarse en sus estudios.

Para el desarrollo de este trabajo la División de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa proporcionó información relativa a las *características individuales, a los antecedentes educativos* y a la trayectoria escolar de los estudiantes que ingresaron en 1995, durante los doce trimestres que establece la universidad como plazo regular para el término de los estudios.

Este trabajo está dividido en cinco capítulos. A continuación se describe brevemente el contenido de cada uno de ellos:

Capítulo 1: Marco teórico del estudio de la deserción.

En este capítulo se desarrolla el marco teórico del estudio de la deserción, elaborado a partir de la revisión de los documentos más importantes que se han producido en este rubro, subrayando la importancia de las investigaciones realizadas por el estadounidense Vincent Tinto.

Capítulo 2: El caso de la Universidad Autónoma Metropolitana, Unidad Iztapalapa, División de Ciencias Básicas e Ingeniería.

Comprende el marco contextual de este trabajo, así como la propuesta de cómo se abordará el concepto de deserción estudiantil bajo este marco. También se incluye el protocolo de investigación donde se plantean los lineamentos generales de investigación que respaldan este trabajo, los objetivos generales y los específicos que subyacen a la hipótesis anteriormente planteada y la justificación del mismo Así mismo, incorpora el sustento teórico de las técnicas utilizadas para el análisis: Análisis de Conglomerados y Análisis de Sobrevivencia. Específicamente, en el Análisis de Conglomerados se utilizó el método no jerárquico de *k medias* y, en el Análisis de Sobrevivencia, el método no paramétrico producto-límite de Kaplan y Meier y el modelo de funciones de riesgos proporcionales de Cox.

Capítulo 3: Análisis de datos.

En este capítulo se lleva a cabo el análisis de datos el cual está dividido en tres etapas fundamentales. La primera, se refiere al análisis preliminar a través del desempeño académico por carrera de los estudiantes hasta el doceavo trimestre. La segunda etapa se refiere al análisis de conglomerados por medio del método no jerárquico de k medias y, la *última etapa se cumple con el análisis de sobrevivencia.*

Capítulo 4: Discusión.

En el capítulo cuarto se discuten los resultados, alcances y limitaciones del trabajo.

Capítulo 5: Conclusiones.

El desarrollo de este trabajo permitió comprobar que las características individuales del estudiante influyen en su probabilidad de éxito en sus estudios superiores. De hecho, se encontró que la edad al ingreso, la carrera y el género son los factores que más influyen en que un estudiante abandone sus estudios. Además, se encontró que las mujeres tienen mayor probabilidad de terminar los estudios que los hombres; que las carreras cuyos alumnos tienen mayor probabilidad de éxito en sus estudios universitarios son Ingeniería en Computación e Ingeniería Electrónica y que la edad al ingreso óptima para maximizar la probabilidad de terminar la carrera no es precisamente 18 ó 19 años como podría pensarse sino alrededor de 22 años

Cabe resaltar que estos resultados son los obtenidos específicamente para la División de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa. Sin embargo, se recomienda realizar este tipo de estudios para las demás divisiones de la Unidad Iztapalapa y para las tres unidades que integran la Institución así como también para otras Universidades.

Agradezco a la Maestra Magdalena Fresán Orozco, profesor-investigador de la Universidad Autónoma Metropolitana, Unidad Xochimilco y asesora de la Secretaría General Ejecutiva de la ANUIES[1], por su generosa participación y asesoría brindadas en la elaboración de esta *investigación Así mismo, a la División de Ciencias Básicas* e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa por haber proporcionado la información utilizada en este estudio.

El disco que acompaña a este trabajo contiene los anexos que incluyen información utilizada durante el desarrollo de la investigación. El paquete estadístico utilizado para el análisis de datos es el SPSS 8.0 (Statistical Programme for Social Sciences)

[1] Asociación Nacional de Universidades e Instituciones de Educación Superior.

CAPÍTULO 1
MARCO TEÓRICO DEL ESTUDIO DE LA DESERCIÓN

1.1. Antecedentes:

El presente trabajo responde a la imperiosa necesidad de realizar estudios sobre deserción estudiantil en el sistema de educación superior en México. Las estadísticas disponibles son elocuentes. Algunos estudios ya publicados señalan como promedio nacional, que de cada 100 alumnos que ingresan a estudios de licenciatura, entre 50 y 60 culminan las materias del plan de estudios cinco años después y de éstos, únicamente 20 se titulan. De los que se titulan, sólo el 10%, esto es 2 estudiantes, culminan sus estudios a la edad estimada como meta deseable de 24 ó 25 años; los restantes, 98 estudiantes, lo hacen entre los 27 y 60 años de edad[1].

Las consecuencias de este masivo y permanente éxodo de las universidades no son triviales, tanto para los alumnos que desertan como para sus instituciones. Respecto a los primeros el tipo de ocupación, la remuneración y otras retribuciones sociales vinculadas a la educación superior están, en gran medida, condicionadas a la obtención de un grado universitario, lo cual no quiere decir que los individuos que han fracasado en su intento de alcanzar un grado académico no se han beneficiado de la educación superior. Para las instituciones, este problema afecta sus indicadores de eficiencia y pone en tela de juicio su funcionalidad. Por último, es un problema también para la sociedad que invierte en la formación de recursos humanos, porque los desertores ocupan un lugar que hubiese significado una oportunidad para otro aspirante a la educación superior.

Diversas investigaciones realizadas tanto en México como en otros países permiten comprender las distintas facetas del fenómeno de deserción estudiantil en la educación superior Entre éstas, sobresalen las investigaciones realizadas por el norteamericano Vincent Tinto, cuyos resultados y propuestas constituyen el fundamento teórico de la mayoría de los estudios sobre este problema. En el siguiente apartado titulado "teoría de la deserción estudiantil", se analizan las ideas centrales de este autor que constituyeron el sustento principal para este trabajo.

[1] Díaz de Cosío, Roger, "Los desafíos de la educación superior mexicana" en Revista de la Educacion Superior, No 106, abril-junio de 1998, ANUIES, p.8.

1.1.1. Teoría de la deserción en la Educación Superior

Vincent Tinto, investigador norteamericano, desarrolló un modelo teórico para explicar el fenómeno de la deserción a partir de las diferentes teorías y modelos anteriores Su modelo[2], representado en el Diagrama 1, considera la deserción:

> *"como un proceso longitudinal de interacciones entre el individuo y los sistemas académico y social de la universidad, durante el cual las experiencias del estudiante en ambos entornos (ponderados según su integración normativa y estructural) modifican continuamente sus metas y compromisos institucionales y lo conducen a la persistencia en los estudios o a distintas formas de abandono de los mismos."*[3]

El modelo de Tinto tiene como principio la teoría sobre el suicidio, del psicólogo francés Émile Durkheim[4]. Esta teoría establece que la probabilidad de que un individuo se suicide aumenta cuando carece de dos tipos de integración: la que se vincula con los valores morales y la que se refiere a la afiliación a la comunidad La primera se puede interpretar como consecuencia de la incompatibilidad entre los valores del individuo y la sociedad; la segunda, como falta de comunicación e interacción entre el individuo y los restantes miembros de la comunidad.

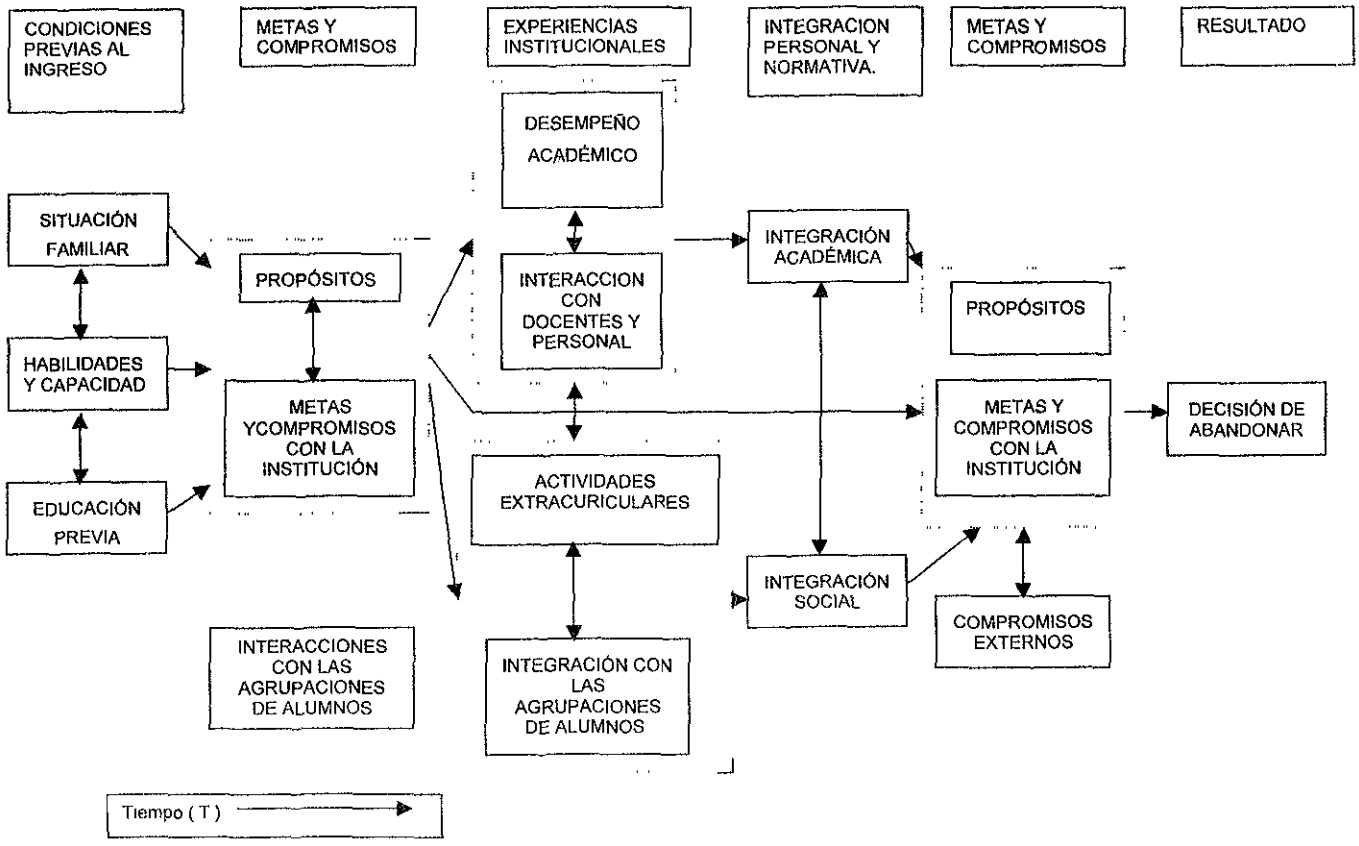
Tinto considera que las condiciones sociales que llevan al suicidio se asemejan a las que conducen a la deserción de la universidad al ser ésta, también, un sistema social. De este modo, la insuficiente interacción y comunicación entre el estudiante y los demás miembros de la comunidad universitaria y la falta de congruencia con los valores existentes en la comunidad aumentan la probabilidad de que un estudiante deserte

La Teoría de Durkheim resulta un modelo sugerente para precisar las condiciones en las cuales se producen las distintas modalidades de abandono de los estudios. Sin embargo, por sí misma, la explicación del fenómeno del suicidio es considerada insuficiente por Vincent Tinto, ya que no toma en cuenta otras características individuales, tales como: sexo, capacidad, origen social, nivel económico, estudios anteriores y desempeño en el examen de admisión, entre otras, que podrían ser fundamentales para explicar este problema.

---

[2] Tinto, Vincent. El abandono de los estudios superiores. Una nueva perspectiva de las causas del abandono y su tratamiento. México, UNAM-ANUIES, 1987 pp 93-136.
[3] *Ibídem*, p 7.
[4] *Ibídem*, pp.107-112.

Diagrama 1. Modelo de abandono institucional



La Teoría de Durkheim resulta un modelo sugerente para precisar las condiciones en las cuales se producen las distintas modalidades de abandono de los estudios. Sin embargo, por sí misma, la explicación del fenómeno del suicidio es considerada insuficiente por Vincent Tinto, ya que no toma en cuenta otras características individuales, tales como· sexo, capacidad, origen social, nivel económico, estudios anteriores y desempeño en el examen de admisión, entre otras, que podrían ser fundamentales para explicar este problema.

Los modelos de deserción, según este investigador, no sólo deben incluir los antecedentes de los estudiantes, sino también las cualidades individuales vinculadas con las expectativas y motivaciones. En cuanto a las expectativas, el modelo de Tinto incluye el compromiso personal del estudiante con las metas académicas y respecto a las motivaciones, el compromiso del individuo con la universidad a donde asiste.

Por otro lado, para atender la naturaleza longitudinal de la deserción estudiantil, Vincent Tinto se basa en los trabajos realizados por el antropólogo holandés Arnold Van Gennep quien sostiene en su estudio titulado *Los ritos de transición* [5] que el proceso de transmisión de las relaciones entre grupos encadenados temporalmente, está determinado por tres distintas fases o etapas, cada una con ceremonias y rituales específicos. Estas fases son: la separación, la transición y la incorporación Cada etapa sirve para trasladar a los individuos de una participación juvenil, a una membresía plena de una sociedad adulta.

La primera etapa, de la separación, se caracteriza por una paulatina declinación en las interacciones con los miembros del grupo de donde proviene la persona. La segunda fase o etapa de transición implica un periodo en el cual la persona comienza a interactuar de nuevas formas con los miembros del nuevo grupo al que se pretende ingresar. Es en esta etapa transicional cuando el individuo adquiere el conocimiento y las habilidades requeridas para la ejecución de su rol específico en el nuevo grupo descartando formas de comportamiento anteriores La tercera fase, la de incorporación, implica asumir nuevos patrones de interacción con miembros del nuevo grupo, y lograr el establecimiento de una membresía competente en ese grupo, en cuanto participante del mismo.

Vincent Tinto[6] afirma que los ritos de transición de Van Gennep constituyen un instrumento de análisis para estudiar el proceso longitudinal de la persistencia estudiantil en la universidad y, por extensión, el del abandono, lo que en particular permite suponer que el proceso que caracteriza a la persistencia es funcionalmente similar al que conduce a la integración en la vida de las

---

[5] Van Gennep, A The Rites of passage, trans By M Viedon and G Caffee Chicago University University of Chicago Press 1960

comunidades humanas en general. Asimismo, está marcado por periodos de transición similares que los individuos deben transitar para continuar en la institución. Además, sugiere que el proceso de deserción estudiantil refleja las dificultades que los alumnos necesitan superar para recorrer con éxito esos periodos de transición.

Tinto considera que, a pesar de que el trabajo de Van Gennep ha permitido avanzar en el desarrollo de una teoría sobre el abandono escolar, no proporciona un medio para analizar los procesos de interacción entre los individuos, en gran medida informales, que llevan a la integración en la vida universitaria.

Algunos autores están en desacuerdo con el modelo de deserción de Tinto. Entre ellos se encuentran el investigador norteamericano John M. Braxton[7], quien revisó el modelo exhaustivamente y concluyó que algunos aspectos carecen de consistencia empírica interna y por tanto que existen conclusiones insuficientemente sustentadas. Sin embargo, Braxton recomienda no abandonar el modelo sino examinarlo a profundidad.

Características importantes en el estudio de la deserción.

Hay una importante cantidad de factores que han sido vinculados al abandono de los estudios. A continuación se reseñan las variables que se implican con mayor frecuencia en este fenómeno.

Diferentes investigadores[8] (Chaín: 1995; Carrillo:1993, Martínez. 1996; Vázquez 1989, Muñiz, 1997) coinciden en que algunas particularidades de los individuos se relacionan con la deserción, entre ellas, las más importantes se refieren a las características familiares, las personales propiamente dichas, los antecedentes de estudios previos y las expectativas acerca de futuros logros académicos.

---

[6] Tinto (1987) *op. cit.* p. 99-106

[7] Braxton, John, *et al*, "Tinto´s separation stage and its influence on first-semester college student persistence" en *Research in Higher education.* Volume 41, No. 2, April 2000.

[8] Entre ellos: Chaín, Ragueb. Estudiantes universitarios· Trayectorias escolares. Xalapa, UV-UAA, 1995, pp. 119-196 , Carrillo Flores, Irma "El abandono escolar en educación superior; ¿una decisión racional o un efecto multifactorial involuntario? Universidad Autónoma de Aguascalientes" en Eficiencia terminal y Calidad Académica en las IES Universidad de Guadalajara. 1993; Martínez Rizo, Felipe "La información sobre los alumnos en las IES", en Propuesta Metodológica para el estudio de los fenómenos de Deserción, Rezago y Eficiencia Terminal en las IES. ANUIES 2001 (en proceso de revisión editorial); Vázquez Chagoyán, Ricardo "La influencia de los estilos cognoscitivos en el rendimiento escolar", en Tinto, Vincent, *et al.* Trayectoria escolar en la educación superior. México, SEP-ANUIES, 1989 pp 113-141; Muñiz Martelón, Patricia "Trayectorias educativas y deserción universitaria en los ochenta", en *Temas de Hoy en la Educación Superior*, Núm 19, ANUIES, 1997, 190 pp.

1 Antecedentes familiares

Las características familiares están correlacionadas con la probabilidad de que un estudiante abandone la universidad. El nivel socioeconómico de la familia se comporta en razón inversa con la deserción; en particular, el investigador mexicano Ragueb Chain destacó que entre los estudiantes provenientes de familias de situación socioeconómica más baja se observan índices más altos de deserción que entre los que pertenecen a familias con una situación socioeconómica más elevada A su vez, Felipe Martínez Rizo[9], en investigaciones realizadas en la Universidad Autónoma de Aguascalientes, confirma que los antecedentes socioeconómicos y el entorno del estudiante no deben minimizarse y, por tanto, celebra la compilación de información socioeconómica por parte de organismos evaluadores como el CENEVAL[10], que permitirán en pocos años profundizar en esta relación.

La probabilidad de permanencia de un estudiante en el sistema educativo superior es mayor en aquellos cuyos padres tienen un mayor nivel educativo[11]. Por ejemplo, en el caso de la Universidad Veracruzana, Ragueb Chain encontró que el porcentaje de padres con licenciatura es el doble en el caso de los estudiantes con estudios regulares, que aprueban sus asignaturas a través de exámenes ordinarios y obtienen calificaciones altas[12].

Tinto señala que otros factores vinculados con los antecedentes familiares influyen también en los logros educativos del estudiante y su desempeño académico Los más importantes de esos factores son la calidad de las relaciones dentro de la familia, el interés y las expectativas de los padres acerca de la educación de sus hijos[13].

2. Características individuales.

Como se mencionó en el apartado anterior, las características familiares representan un factor importante en el desempeño académico de los estudiantes; sin embargo, Vázquez[14] sostiene que la capacidad del sujeto es uno de los elementos decisivos para la persistencia en los estudios.

---

[9] Martínez Rizo, Felipe " La información sobre los alumnos en las IES" en Propuesta Metodológica para el estudio de los fenómeno de Deserción, Rezago y Eficiencia Terminal en las IES. ANUIES 2001 (en proceso de revisión editorial)

[10] Centro Nacional de Evaluación

[11] "El factor social más general que determina las diferencias de rendimiento escolar entre los individuos es el origen de clase, su pertenencia a un estrato social.[. .] dado que la familia es el núcleo de socialización primaria los aspectos diversos de la socialización diferencial en ese periodo podrían dar por resultado desarrollo (subdesarrollo) diferenciales que se manifiestan en la exiguidad de sus progresos ulteriores tanto en la escuela como en la vida en general." Vázquez Chagoyan, Ricardo "La influencia de los estilos cognoscitivos en el rendimiento escolar" en Tinto, Vincent, *et al* Trayectoria escolar en la educación superior. México, SEP-ANUIES, 1989. pp 114-116

[12] Chain, Ragueb, *op cit*, p. 137.

[13] Tinto, Vincent, *et al* Trayectoria escolar ... pp 14-15.

[14] Vázquez C . Ricardo "La influencia de los estilos cognoscitivos en el rendimiento escolar" en Tinto, Vincent, *et al*, Trayectoria escolar ., pp 119-120.

Indudablemente existe vinculación entre la capacidad del estudiante y su permanencia en la universidad. Con base en sus investigaciones, Tinto considera que la capacidad del alumno es cuantificable por medio de las calificaciones obtenidas en la etapa educativa anterior a los estudios universitarios. Chaín[15] señala, además, que los alumnos con estudios regulares, exámenes ordinarios y calificaciones altas tuvieron, en general, un mejor desempeño en la educación media superior que los alumnos que no presentan estas características.

Por otro lado, Tinto[16] destaca, aunque no con la misma relevancia que la capacidad individual, las diferencias entre la personalidad de los estudiantes desertores y de aquellos que persisten en los estudios. Señala que los desertores tienden a ser más impulsivos que quienes persisten en los estudios; además carecen de un sólido compromiso emocional con la educación y son incapaces de sacar provecho de las experiencias pasadas en la misma medida que los estudiantes que prosiguen sus estudios.

La relación del género con el éxito en los estudios, también ha sido preocupación de los estudiosos interesados en el fenómeno de la deserción Por ejemplo, el profesor de la Universidad Pedagógica Nacional, Ricardo Vázquez[17], ha observado que, en general, las mujeres tienen mejor rendimiento que los hombres, y lo atribuye a dos posibles causas: a) las mujeres poseen una inteligencia verbal más desarrollada que los varones y/o b) la educación familiar tradicional de las mujeres responde a las expectativas de los profesores respecto al concepto "buen alumno" dado *su carácter pasivo y su receptividad.*

3. Antecedentes educativos

El comportamiento en el nivel medio de enseñanza, evaluado ya sea mediante el promedio de calificaciones o por algún otro mecanismo, se ha reportado con frecuencia como un indicador importante del futuro desempeño del estudiante Otros factores determinantes en los resultados que logra un estudiante son aquellos relacionados con la calidad de la escuela de nivel medio, tales como sus instalaciones y su personal docente. Vincent Tinto sugiere que este factor también puede afectar el desempeño estudiantil y, por consiguiente, su permanencia en la institución. En la Universidad Veracruzana, Ragueb Chain investigó exhaustivamente la relación entre la trayectoria escolar universitaria y el tipo de bachillerato de procedencia; encontró que los estudiantes regulares, con exámenes ordinarios y calificaciones altas provienen de bachilleratos más eficientes (aquellos con mayor porcentaje de alumnos admitidos en la universidad)[18]

---

[15] Chain, Ragueb, *op cit*, p 132.
[16] Tinto, Vincent, *et. al.* Trayectoria escolar , pp.15-16.
[17] Vázquez C.,Ricardo. *op cit*, pp. 120-121.

4. Compromiso con la meta

Tinto se refiere a una variable individual muy peculiar: el compromiso con la meta. Es decir, el compromiso de completar exitosamente los estudios y considera que éste es el factor que más influye en el desempeño académico, después de la capacidad personal[19]

El compromiso con la meta se incluye junto con los antecedentes familiares y las experiencias previas en el modelo de deserción de Tinto (Véase diagrama 1). Este planteamiento considera a los compromisos de un estudiante universitario como el reflejo de un proceso multidimensional de interacciones entre el individuo, su familia y las experiencias anteriores en la escuela. La representación de esos antecedentes, en particular los que conciernen a la familia, en la continuidad de los estudios, se produce en gran parte por conducto de la influencia de dichos antecedentes sobre el desarrollo de los compromisos educativos e institucionales del individuo En México, en particular en la Universidad Autónoma de Aguascalientes, Felipe Martínez Rizo[20] ha realizado trabajos en donde pretende medir el compromiso con la meta a través de las variables subjetivas: actitudes de logro, interés por los estudios, interés por la carrera, interés por la institución y coincidencia con la filosofía institucional. Esta es una dimensión interesante, sin embargo, la indagación de estas variables no constituye una constante en los bancos de información sobre las trayectorias escolares y difícilmente puede ser incluida en un estudio estadístico o probabilístico orientado a sustentar la toma de decisiones para mejorar la persistencia y abatir el abandono en el nivel superior

5. Interacción en el medio universitario

Vincent Tinto supone que si los factores externos al proceso educativo tales como la oferta y demanda en el mercado laboral o la existencia de políticas restrictivas (como cualquier tipo de discriminación), permanecieran constantes en una sociedad, la deserción sería entonces, consecuencia exclusivamente de las experiencias personales en los sistemas académico y social de la universidad[21].

En el diagrama 1 se muestra cómo los compromisos educativos e institucionales se ubican al principio y al final del modelo ya que funcionan como variables básicas (de entrada) como de proceso y proporcionan el componente dinámico de la progresión del individuo en el sistema educativo

---

[18] Chain, Ragueb, *op cit*, pp. 124-131.
[19] Tinto, Vincent, Trayectoria escolar..., pp 17-19.
[20] Martínez R., Felipe "Diseño de investigación para el estudio de la deserción Enfoque cuantitativo transversal " En Tinto, Vincent, *et al.* Trayectoria escolar..., p 288.
[21] Tinto, Vincent, *et al.* Trayectoria escolar , p 19.

6. Integración académica

La integración académica de un estudiante se puede medir a través de sus calificaciones y de su desarrollo intelectual durante los años que estudia en la universidad Tinto afirma que.

> *"...ambos aspectos implican elementos estructurales y normativos, el primero se relaciona más directamente con el cumplimiento de determinados criterios explícitos del sistema académico, mientras que el último atañe más a la identificación del individuo con las normas de ese sistema."*[22]

Las calificaciones, de acuerdo con algunos autores, representan la forma más concreta de recompensa en el sistema académico universitario. Son una forma de premiar la participación del estudiante en la institución, que puede ser utilizada como recurso tangible para su futura movilidad educativa y profesional. Por otra parte, el desarrollo intelectual representa un modo de recompensa más subjetivo; puede interpretarse como la evaluación del sistema académico por el individuo Las calificaciones expresan tanto la capacidad personal como las preferencias institucionales por determinar estilos de comportamiento académico. Ragueb Chain utiliza las calificaciones del estudiante como un criterio de clasificación para cuantificar la trayectoria escolar y determinar la adaptación de los alumnos a la institución.[23]

7 Integración social

Las decisiones del estudiante, vinculadas con la permanencia en la universidad, partiendo de grados previos de compromiso con las metas educativas y la institución, también pueden ser afectadas por la integración del sujeto en el sistema social de la institución Ragueb Chain confirmando esta afirmación, incluyó en su estudio realizado en la Universidad Veracruzana, una variable (relaciones con los compañeros) que pretendía medir la integración social del estudiante. Encontró que una tercera parte de los estudiantes tiene una relación de camaradería, la mitad una de cooperación, el 7% de indiferencia y el 5% de competencia[24]. Posiblemente, la integración adecuada a los círculos sociales universitarios de un alumno incremente las probabilidades de su permanencia en la institución.

---

[22] *Ibídem*, pp. 19-20.
[23] Chain, Ragueb, *op cit*, p. 93.
[24] *Ibídem*, p 184

1.1.2. Modelo para el estudio de la deserción.

Como ya se mencionó, el estado de alumno desertor depende de varios factores tales como:

1. Antecedentes familiares.
2. Características individuales.
3. Antecedentes educativos
4. Compromiso con la meta.
5. Integración con el medio Universitario.
6. Integración académica.
7. Integración social.

Estos factores sugieren el análisis del problema mediante un modelo multivariado en el que la respuesta en estudio se puede explicar con variables relativas a las siete dimensiones anteriores Sin embargo, actualmente la falta de información disponible en forma sistemática por parte de las Instituciones de Educación Superior (IES) no permite la construcción de dicho modelo. La División de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa solamente proporcionó para este estudio información relativa a las dimensiones de Características individuales y Antecedentes educativos. Aunque al ingreso, el alumno contesta un cuestionario con información relacionada con sus antecedentes familiares, no se pudo tener acceso a ella debido a la falta de infraestructura apropiada en 1995, por parte de la universidad. Por tanto, los alcances de este estudio se limitan a la siguiente información:

- Género
- Edad al ingreso
- Escuela de procedencia del nivel educativo anterior
- Promedio en el nivel educativo anterior
- Carrera
- Periodo de ingreso a la universidad
- Resultado en el examen de admisión en las áreas de razonamiento verbal, razonamiento específico y conocimientos generales.

1.2. Objetivos

1.2.1. Objetivo general

Construir un modelo de análisis y diagnóstico de la trayectoria escolar de los estudiantes, a través de las diversas etapas que comprenden los planes de estudio de Licenciatura de las IES públicas con el fin de detectar oportunamente estudiantes con alto riesgo de abandono de sus estudios.

1.2.2. Objetivos específicos.

1.2 2 1. Construir un modelo de análisis y diagnóstico de la trayectoria escolar de los estudiantes que permita estimar la probabilidad de que un estudiante no abandone la universidad en las diversas etapas que comprenden los planes de estudio de una área del conocimiento. Específicamente, estudiantes de la cohorte de ingreso 1995 de las 9 diferentes carreras que ofrece la división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa.

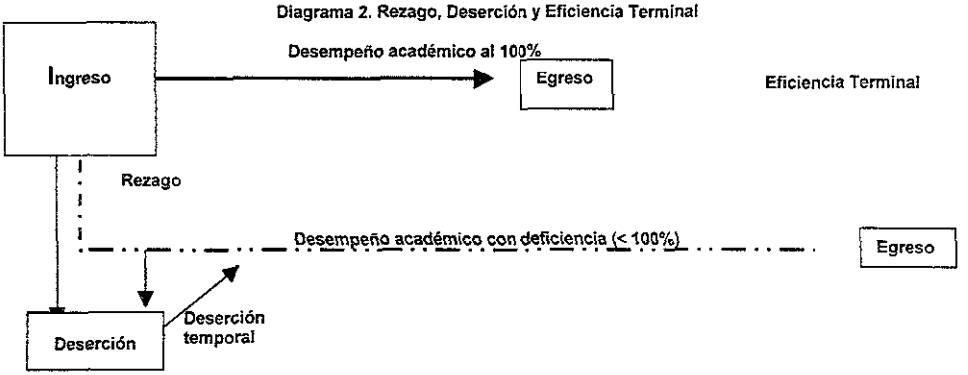
1.2.2.2. Elucidar las características de los estudiantes, de la cohorte de ingreso 1995, que influyen con mayor impacto en la permanencia o el abandono de los estudios en el nivel universitario, en las distintas carreras que ofrece la división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa.

1.3. Justificación del estudio

La *deserción estudiantil*, definida[25] como el abandono definitivo o temporal de los estudios superiores, es un problema al que se enfrentan las IES y que resulta un factor determinante en la calidad de vida de la sociedad. Es necesario, por tanto, realizar investigaciones que analicen y evalúen este fenómeno con el objeto de implantar medidas preventivas que contribuyan a abatir este problema y, de tal forma, contribuir a la funcionalidad de la educación superior.

El *rezago estudiantil*,[26] entendido como el atraso o retraso de los estudiantes en la inscripción a las asignaturas, según la secuencia establecida en el plan de estudios y la *eficiencia terminal*, definida como[27] la relación cuantitativa entre los alumnos que ingresan y los que egresan de una misma cohorte, son consideradas junto con la *deserción* tres facetas de un mismo fenómeno frecuente al nivel de la Licenciatura. La naturaleza de estos tres aspectos sugiere su análisis conjunto, tomando como eje central la deserción, ya que el rezago se considera como una de sus causas y la eficiencia terminal como su consecuencia institucional. El siguiente diagrama 2 describe dicho proceso:

Diagrama 2. Rezago, Deserción y Eficiencia Terminal



El desempeño académico se refiere al avance de créditos del estudiante durante su trayectoria escolar, que puede ser igual o menor al establecido por el plan de estudios de la universidad. Si este avance es menor, entonces se entiende que el estudiante es un alumno rezagado[28], es decir, un alumno que por cualquier razón, no mantiene el ritmo regular del plan de estudios y su egreso ocurre en una fecha posterior a la establecida en dicho plan

[25] Tinto, Vincent El Abandono de los estudios superiores· Una nueva perspectiva de las causas del abandono y su tratamiento UNAM y ANUIES 1992.
[26] Altamira, Rodríguez. El análisis de las trayectorias escolares como herramienta de evaluación de la actividad académica universitaria 1997.
[27] Tinto, Vincent *op. cit* El Abandono de los estudios superiores Pág. 297
[28] Rangel, Alfonso. Glosario de la Educación Superior ANUIES. México 1996

La deserción puede darse ya sea en los alumnos con un desempeño académico del 100% o con un desempeño académico con deficiencia, es decir, en alumnos rezagados

Al describir la deserción se distingue entre el abandono de una determinada universidad (abandono institucional) y el que se refiere a todo el sistema de educación superior (abandono del sistema). Estas situaciones son diferentes no sólo en su naturaleza sino en su variabilidad con la que se efectúan en los distintos grupos de la población estudiantil de la universidad. No todos los estudiantes que abandonan un Institución quedan fuera del sistema de Educación Superior. Muchas deserciones institucionales son, en realidad, movimientos migratorios de alumnos a otras universidades del sistema -transferencia institucional- y otras resultan sólo una interrupción temporal de los estudios formales -desertores temporales.

La deserción puede presentarse en las siguientes modalidades[29]:

- Abandono o suspensión voluntaria y definitiva de los estudios y del sistema general de educación superior.
- Baja atribuible a deficiencias académicas y bajo rendimiento
- Cambio de carrera en la misma u otra institución
- Baja por motivos reglamentarios, de índole no académica

Según Vincent Tinto[30] la probabilidad de completar exitosamente la formación de licenciatura en el tiempo estipulado en los planes de estudio, está determinada por diversos factores tales como la edad, el sexo del estudiante, el campo de estudio, la integración al medio universitario y otros factores relacionados con los antecedentes académicos del estudiante, como su condición socioeconómica y el nivel cultural de su familia. Por otro lado, diversos investigadores, al igual que Ragueb Chaín[31] en la Universidad Veracruzana, han estudiado la capacidad del examen de selección para predecir el comportamiento del alumno durante el primer año de sus estudios (rendimiento académico durante el primer año de los estudios superiores). Sin embargo, en la mayoría de las investigaciones sobre el problema de la eficiencia terminal de los programas de estudio en el nivel superior se trabaja con modelos univariados, sin haber conseguido hasta este momento un modelo multivariado de predicción, aplicable a la realidad nacional. Por ello se hace necesaria la búsqueda de estrategias que, a partir de la identificación de estas variables, permitan modelar estadísticamente este fenómeno

[29] Tinto, Vincent. *op. cit* El Abandono de los estudios superiores..
[30] *Ibidem.*
[31] Chain, Ragueb (1995) *Op. Cit.*

Resulta de particular interés la aplicación de técnicas estadísticas en los estudios de deserción con la potencia necesaria para estimar el riesgo que tiene un estudiante, dadas ciertas características, de abandonar sus estudios de licenciatura.

El presente trabajo pretende estimar la probabilidad de que el alumno no abandone sus estudios en alguno de los doce trimestres que la universidad establece como periodo regular para el termino de los estudios.

Para la realización de este estudio, se cuenta con información del avance en créditos durante los 12 trimestres que establece la universidad como plazo regular, de los estudiantes que ingresaron a la División de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa, en la primavera y otoño de 1995. Además, se dispone por trimestre, del número de materias inscritas y de éstas, el número de aprobadas o reprobadas y el promedio obtenido

Respecto a la información general de los estudiantes se dispone de la edad, promedio en el nivel educativo anterior, escuela de procedencia del nivel educativo anterior y el porcentaje de respuestas correctas en las distintas áreas del examen de admisión, es decir, razonamiento verbal, razonamiento matemático y conocimientos específicos.

1.4. Hipótesis

La permanencia de los estudiantes en los programas de licenciatura está determinada por la edad, el género, la carrera, el período de ingreso, factores (relacionados con los antecedentes académicos del estudiante) como el promedio en el nivel educativo anterior y la escuela de procedencia del mismo nivel, así como por el resultado obtenido en el examen de admisión. Por tanto, con esa información es posible determinar el riesgo que tiene un estudiante de abandonar o rezagarse en sus estudios.

CAPÍTULO 2
EL CASO DE LA UNIVERSIDAD AUTÓNOMA METROPOLITANA, UNIDAD IZTAPALAPA DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

2.1. Marco Contextual.

2.1.1 Universidad Autónoma Metropolitana[1]

La Universidad Autónoma Metropolitana es una institución pública y autónoma, cuyo modelo educativo está sustentado en el modelo departamental. Imparte 59 licenciaturas y 60 posgrados en el ámbito nacional y 7 internacionales, en las áreas de Ciencias Básicas e Ingeniería, Ciencias Biológicas y de la Salud, Ciencias Sociales y Humanidades y Ciencias y Artes para el Diseño. El periodo escolar está estructurado en forma trimestral para ambos niveles.

Al trimestre de Otoño 96, la matrícula escolar en licenciatura y posgrado estaba conformada por 45,000 alumnos.

Otra característica de esta Institución educativa es la figura de profesor – investigador. La planta académica está formada por 2,315 profesores - investigadores con dedicación de tiempo completo, por 477 de medio tiempo y por 301 de tiempo parcial. Esto significa que cerca del 75% del personal académico de la universidad dedica tiempo completo a sus labores de docencia e investigación.

El número de profesores - investigadores de tiempo completo que pertenece al Sistema Nacional de Investigadores se ha incrementado significativamente durante los últimos 12 años. En 1984 era de sólo 45; para 1997 el número de ellos aumentó a 410.

La UAM se integra por una Rectoría General y tres Unidades Universitarias: Unidad Azcapotzalco, Unidad Iztapalapa y Unidad Xochimilco.

Las Unidades Universitarias tienen bajo su responsabilidad el desarrollo de las actividades relacionadas con la impartición de los programas docentes de nivel licenciatura y de posgrado, la realización y evaluación permanente de los programas y proyectos de investigación, la formación integral de los estudiantes, la generación de acciones tendientes a difundir el conocimiento, la extensión de la cultura y la vinculación con el entorno. Cada una de ellas cuenta con sus propios órganos académicos y administrativos para impulsar y desarrollar sus actividades

2.1.2. Reglamento de Estudios Superiores[2]

A continuación se enuncian algunas partes de artículos incluidos en el Reglamento de Estudios Superiores que son de utilidad para este trabajo

- Articulo 14

Adquirirán la calidad de alumno en los estudios de licenciatura y de posgrado, con todos los derechos y obligaciones que establezcan las normas y disposiciones reglamentarias de la universidad, quienes cumplan con los requisitos de ingreso, hayan sido aceptados por la universidad y realicen oportunamente los trámites de inscripción

- Articulo 18

La calidad de alumno o de participante se pierde por las siguientes causas:

I. Por conclusión del plan de estudios;

II. Por renuncia expresa a la universidad o tácita a la inscripción a un año escolar,

III. Por vencimiento del plazo máximo previsto para cursar los estudios;

IV. Por resolución definitiva dictada por el órgano colegiado competente;

VII. Para el nivel de licenciatura, además:

a. Cuando no se hubiere acreditado una misma unidad de enseñanza-aprendizaje mediante cinco evaluaciones globales y de recuperación, y

b. Cuando al cursar el tronco general, el número de evaluaciones globales o de recuperación que no se hubieren acreditado, fuere igual al número de unidades de enseñanza-aprendizaje que lo integran más dos, de acuerdo con el plan de estudios vigente de la licenciatura en la que se está inscrito;

- Articulo 45

En licenciatura, todos los alumnos deberán cubrir la totalidad de los créditos en un plazo que no excederá a diez años, mismo que se computará a partir del primer ingreso a la universidad.

El plazo mínimo para cursar la totalidad de los créditos no podrá ser menor a diez trimestres, en el caso de carreras con una duración prevista de doce, y de trece trimestres en el caso de carreras con una duración prevista de quince.

[2] Dirección electronica: www.uam.mx

- Articulo 47

Quienes hubieren interrumpido sus estudios podrán adquirir nuevamente la calidad de alumno, cuando no hubiere vencido el plazo máximo establecido, pero deberán sujetarse al plan de estudios vigente a la fecha de su reingreso.

En el caso de licenciatura y tratándose de una interrupción mayor de seis trimestres lectivos consecutivos, deberán aprobar un examen de conjunto de las unidades de enseñanza-aprendizaje
acreditadas, según lo establezca el Consejo Divisional correspondiente. El propio Consejo determinará de acuerdo con el resultado del examen, los contenidos del plan de estudios que quedan por acreditar.

- Articulo 48

El interesado en adquirir nuevamente la calidad de alumno de licenciatura deberá:

I. Presentar al Consejo Divisional correspondiente solicitud por escrito debidamente fundamentada.

II. Haber cubierto como mínimo el 75% de los créditos correspondientes a los planes y programas de estudio de licenciaturas de doce trimestres y el 80% de las de quince trimestres.

III Aprobar un examen de conjunto de las unidades de enseñanza-aprendizaje acreditadas, si se hubiesen interrumpido los estudios por más de seis trimestres consecutivos.

IV. Presentar la solicitud dentro de los seis trimestres lectivos contados a partir del vencimiento del plazo máximo

[2] *ibidem*

2.2. Propuesta: El estudio de la retención estudiantil

Para la realización de este estudio, se cuenta con información del avance en créditos durante 12 trimestres, de los estudiantes que ingresaron a la división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa en la primavera y otoño de 1995. Cabe señalar que el plazo que la universidad establece como regular para finalizar las 9 carreras impartidas en esta División es de 12 trimestres. Además, se dispone por trimestre, del número de créditos inscritos y de éstos, el número de créditos aprobados y el promedio obtenido en las materias correspondientes a los créditos aprobados.

La siguiente tabla 1 resume la información proporcionada por la Universidad Autónoma Metropolitana.

Tabla 1. Información general disponible de los estudiantes que ingresaron a la División de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana Unidad Iztapalapa

Nombre de la variable	Descripción de las variables	Escala de medida
SEX	Género del alumno	Nominal
EDAD	Edad en años en el momento del ingreso	Razón
NAL	Nacionalidad	Nominal
REGFED	Registro Federal de Contribuyente	Nominal
ESCUELA	Escuela donde el estudiante obtuvo el certificado de Educación Media Superior	Nominal
PROMEDIO	Promedio obtenido por el estudiante en la Educación Media Superior	Razón
PTO	Porcentaje total de respuestas correctas en el examen de selección	Razón
PRV	Porcentaje total de respuestas correctas en razonamiento verbal en el examen de selección	Razón
PRM	Porcentaje total de respuestas correctas en razonamiento matemático en el examen de selección	Razon
PCO	Porcentaje total de respuestas correctas en conocimientos específicos en el examen de selección	Razón
INGRESO	Trimestre en que el estudiante inició sus estudios (la Universidad Autónoma Metropolitana tiene dos trimestres de ingreso que identifica como primavera y otoño)	Nominal
PLA	Programa de Licenciatura en el cual se encuentra inscrito el estudiante (Carrera)	Nominal
NUMCRE	Número total de créditos establecidos en el plan de estudios	Razón
NUMTRIM	Número de trimestres establecido en el plan de estudios Los 9 planes de estudios de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa abarcan 12 trimestres	Razón

Nombre de la variable	Descripción de las variables.	Escala de medida
NUMINSCi	Número de materias que inscribió en el trimestre i	Razón
CREDAPRi	Créditos aprobados en el trimestre i	Razón
CREDOBLi	Créditos obligatorios en el trimestre i	Razón
PROMINSi	Promedio que obtuvo con las materias que inscribió en el trimestre i	Razón
EDO	Situación escolar del alumno al trimestre 16 1. Activo 2 No reinscrito 3. Baja definitiva 4. Egresado 5. Baja reglamentaria 6 Inscrito en blanco 7. Créditos cubiertos 8 Abandono de más de 6 trimestres Nota: Recordar que el plazo regular de los estudios es de 12 trimestres. La situación escolar de los alumnos al trimestre 16 refleja su estado cuatro trimestres después de que el egreso debió haber sucedido.	Nominal
CREDAPR	Total de créditos del plan de estudios aprobados por el estudiante al doceavo trimestre.	Razón
UTRINSC	Último trimestre en que el alumno aparece registrado como tal en la Dirección de Sistemas Escolares	Ordinal
UTRACT	Último trimestre en que el alumno tiene créditos aprobados o reprobados en los registros de la Dirección de Sistemas Escolares	Ordinal
NA	Número de asignaturas con calificación reprobatoria en el tronco general de asignaturas	Razón

El abandono estudiantil adopta dos formas[3]: la exclusión por razones académicas y la deserción voluntaria.

El Reglamento General de Alumnos[4] de la Universidad Autónoma Metropolitana define las razones académicas por las cuales un estudiante pierde su calidad de alumno. El artículo 45 establece como plazo máximo para finalizar los estudios de Licenciatura 30 trimestres que equivalen a 10 años, contados a partir del período de ingreso. Entonces, si un estudiante prolonga sus estudios por más de 10 años, el sistema lo excluye y es acreedor al término desertor Sin embargo, al contar con las bases de datos provenientes de los estudiantes cuyo ingreso fue en 1995 no es posible conocer la trayectoria académica de los estudiantes por el periodo mencionado

[3] Tinto, Vincent. El abandono de los estudios superiores. Una nueva perspectiva de las causas del abandono y su tratamiento Mexico, UNAM-ANUIES, 1987 pp 93-136.
[4] Universidad Autónoma Metropolitana, Reglamento General de los Alumnos www.uam mx

A su vez, el artículo 47 y 48 establecen que si un alumno interrumpe sus estudios por más de dos años también pierde su calidad de alumno, pero puede recuperarla siempre y cuando tenga presente una evaluación global de los conocimientos adquiridos antes de la interrupción de sus estudios, en un plazo máximo de 6 períodos lectivos, contados desde el vencimiento del segundo año de abandono. En este caso, tampoco es posible determinar los sujetos a los que aplica este artículo, dado que al tercer año de abandono, no se puede determinar si un estudiante presentará la evaluación global en algún momento de los 6 períodos lectivos que la legislación permite.

Dadas las limitaciones expuestas, se consideró conveniente la búsqueda de un término alterno que permitiera abordar el concepto de deserción estudiantil que originalmente se planteó Con la información disponible se puede determinar si un alumno es retenido en la universidad Es decir, si un alumno habiéndose matriculado en un trimestre dado, aparece matriculado en el siguiente Este razonamiento sugiere el término retención *estudiantil.* Es cierto, que bajo este concepto también se enfrentan algunos problemas como los casos de estudiantes que abandonan la universidad solamente por un trimestre. Sin embargo bajo este concepto es posible inferir más certeramente sobre la permanencia de los estudiantes en la universidad

Por tanto, no se utilizará el término *deserción estudiantil* sino el término *retención estudiantil,* bajo el supuesto que una alta retención de la universidad implica una baja tasa de deserción y viceversa.

Para este trabajo un alumno es no - retenido en la universidad si:

- No se inscribe en ningún trimestre posterior al primero.
  Todos los estudiantes de nuevo ingreso son inscritos automáticamente a las asignaturas que el plan de estudios establece para el primer trimestre. En los trimestres consecutivos, los alumnos se inscriben por sí mismos. Cabe señalar que la Universidad Autónoma Metropolitana permite que una vez inscritas las materias, los alumnos en un periodo determinado puedan darlas de baja entrando de este modo a la clasificación de alumno no - retenido.

- No se inscribe por más de 6 periodos consecutivos.
  Bajo el término de deserción, en los casos en los que el abandono sea a partir del séptimo trimestre y hasta el doceavo, se desconoce si el alumno en los dos siguientes años presentará la evaluación que le regresa su carácter de alumno. En general, si el alumno no se inscribe durante 6 periodos consecutivos sabemos que al menos no ha sido retenido por la universidad durante 6 trimestres lo que lo coloca en un estado de potencial desertor.

- El desempeño académico acumulado.
  Un estudiante puede inscribirse en cada uno de los trimestres pero habrá casos en que los créditos acumulados hasta el doceavo trimestre no le permitirían finalizar sus estudios ni en los 10 años establecidos por la universidad En este caso no se puede decir que sea un alumno desertor; sin embargo, se le puede clasificar como un estudiante potencialmente desertor.

A continuación se enuncian las ventajas de utilizar el concepto de retención estudiantil ante el concepto de deserción estudiantil

1. El plazo máximo de 10 años para terminar los estudios pierde relevancia dado que con el término retención, la importancia radica en conocer si el estudiante es retenido en la universidad por determinado periodo de tiempo o si el desempeño académico evidenciado, suponiendo que permanezca en la Institución los 10 años que la legislación universitaria establece, le permitiese aprobar todos los créditos que comprende el plan de estudios que eligió. En el apartado 3.4.3 del capítulo 3 se definen con precisión las condiciones para considerar si un alumno es retenido o no por la universidad.

2 Para los casos en que el estudiante abandona sus estudios a partir del séptimo trimestre y por más de 2 años, la incertidumbre sobre si éste realizará la evaluación global, al cabo de los 6 periodos lectivos establecidos, queda también resuelto por las razones ya expuestas.

2.3. Protocolo de Investigación[5]

2.3.1. Definición del tipo de investigación

El tipo de investigación se elige en función del tipo específico de problema que se quiere abordar y de los recursos de que se dispone.

Los criterios de clasificación de los estudios se definen sobre la base del periodo en que se capta la información, la evolución del fenómeno estudiado, la comparación de poblaciones y la intervención del investigador en el estudio

A continuación se clasifica el tipo de investigación, que se realizará en este trabajo, en función de la información con la que se cuenta y los objetivos mencionados anteriormente.

1 De acuerdo con el periodo en el que se capta la información:

a. Retrospectivo
Estudio del cuál se obtuvo información anteriormente y con fines ajenos a los que se persiguen en la investigación.

Se cuenta con la información de la trayectoria escolar durante 12 trimestres de los estudiantes de la cohorte de ingreso 1995, en el periodo primavera y de la cohorte de ingreso 1995 en el periodo otoño de la división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana – Unidad Iztapalapa

2 De acuerdo con la evolución del fenómeno estudiado.

a Longitudinal
Estudio en el que se miden las variables de interés en diversos momentos de una trayectoria determinada. Implica el estudio de la evolución dichas variables en el tiempo y, por consecuencia su comparación de las distintas etapas consideradas en la investigación

[5] Méndez Ramírez, Ignacio *et. al.* El Protocolo de la Investigación Editorial Trillas México 1984

El desempeño académico de los estudiantes será medido durante los 12 trimestres que establece la universidad como plazo regular[6].

3 De acuerdo con la comparación de las poblaciones.

a. Descriptivo

Estudio que sólo cuenta con una población de la que se tiene como objeto describir en función de un grupo de variables y respecto de la cual no existen hipótesis centrales. En algunos casos se pueden tener hipótesis de relaciones de algunas variables dentro de la misma población.

Se analizará la evolución académica de los estudiantes de la cohorte de ingreso correspondiente al año de 1995 y que debió terminar sus estudios durante 1999, con el objeto de encontrar las características que determinan con mayor fuerza su trayectoria escolar.

b. Comparativo

Estudio en el cual existen dos o más poblaciones y se quieren comparar algunas variables para contrastar una o varias hipótesis centrales.

El estudio es comparativo dado que pretende:

I. Comparar la trayectoria escolar de los estudiantes de las 9 carreras que ofrece la división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa.

II. Comparar las características de los estudiantes retenidos con los no – retenidos en la Universidad.

En lo que toca a la forma de abordar el fenómeno.

- Relación efecto-causa: Se parte de dos o más grupos de unidades de estudio que presentan cierto fenómeno considerado como efecto en varias modalidades Se

[6] Se entiende como plazo regular el plazo que la División de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa establece para el estudio de los 9 diferentes planes de estudio

retrocede al pasado para determinar el factor causal y la proporción en que se presentó en los diferentes grupos

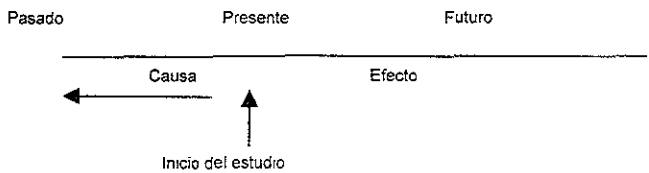


Diagrama 1: Estudio retrospectivo de efecto a causa.

Por tanto, de acuerdo con lo expuesto anteriormente, éste es un estudio longitudinal, retrospectivo, descriptivo y comparativo, que será abordado con relación efecto-causa

2.3.2. Definición de la población objetivo

Cohorte es el grupo de individuos que tuvieron alguna experiencia en común o que comparten alguna característica en específico. Entonces, la población objetivo es la cohorte de alumnos que ingresaron a la división de Ciencias Básicas e Ingeniería (CBI) de la Universidad Autónoma Metropolitana, Unidad Iztapalapa en 1995. Por tanto, las unidades de estudio son los 785 alumnos que ingresaron a la división de CBI de la Universidad Autónoma Metropolitana, Unidad Iztapalapa en 1995.

La población objetivo está dividida en periodo de ingreso y carrera.

2.3.2.1. Características generales:

La característica general de la población objetivo es que las unidades de estudio ingresaron en 1995 a la división de CBI de la Universidad Autónoma Metropolitana, Unidad Iztapalapa.

2.3.2.2. Ubicación temporal y espacial

Para la población objetivo se cuenta con información de la trayectoria escolar de los estudiantes que la conforman durante los 12 trimestres que establece la universidad, como plazo regular para finalizar los estudios de Licenciatura. Entonces, la unidad de tiempo es el trimestre.

La ubicación espacial de la población objetivo es la división de CBI en la Universidad Autónoma Metropolitana – Iztapalapa.

2.3.3. Manejo estadístico de la información:

Se cuenta con todas las unidades de estudio que forman la población objetivo, es decir, no fue necesario muestrear.

2.3.4. Especificación de las variables:

Para la población objetivo se tienen las siguientes variables potencialmente predictivas

VARIABLES POTENCIALMENTE PREDICTIVAS		
Variable	**Descripción de las variables.**	**Escala de medida**
SEX	Género del alumno	Nominal
EDAD	Edad en años en el momento del ingreso	Razón
PROMEDIO	Promedio obtenido por el estudiante en la Educación Media Superior	Razón
ESCUELA	Escuela donde el estudiante obtuvo el certificado de Educación Media Superior	Nominal
PTO	Porcentaje total de respuestas correctas en el examen de selección	Razón
PRV	Porcentaje total de respuestas correctas en razonamiento verbal en el examen de selección	Razón
PRM	Porcentaje total de respuestas correctas en razonamiento matemático en el examen de selección	Razón
PCO	Porcentaje total de respuestas correctas en conocimientos generales en el examen de selección	Razón

VARIABLES PARA FORMAR LAS TRAYECTORIAS ESCOLARES DE LOS ALUMNOS		
Variable	Descripción de las variables	Escala de medida
PAPRINSi	Porcentaje de materias que aprobó con respecto a las inscritas en el trimestre i	Razón
PROMINSi	Promedio que obtuvo con las materias que inscribió en el trimestre i	Razón
NUMINSCi	Número de materias que inscribio en el trimestre i	Razón
PAPRINSi	Porcentaje de materias que aprobo con respecto a las inscritas en el trimestre i	Razón
PROMINSi	Promedio que obtuvo con las materias que inscribió en el trimestre i	Razón
NUMINSCi	Número de materias que inscribió en el trimestre i.	Razón
CREDAPRi	Créditos aprobados en el trimestre i	Razón
CREREPRi	Créditos reprobados (o no aprobados) en el trimestre i	Razón
CREDOBLi	Créditos obligatorios en el trimestre i.	Razón
PAPROBLi	Porcentaje de créditos aprobados respecto a los obligatorios en el trimestre i.	Razón

2.4. Técnicas para el análisis.

2.4.1. Análisis de sobrevivencia.

2.4.1.1. ¿Qué es el análisis de sobrevivencia?

Dada una variable de interés cuyos valores son registrados durante el tiempo, en un conjunto de individuos, hasta que ocurre un determinado evento de interés que está en función de la variable medida, el objetivo del análisis para este tipo de información registrada es estimar, en función del tiempo, la probabilidad de que ocurra dicho evento de interés.

Supongamos que en un hospital se desea estimar la probabilidad, para los N pacientes con úlcera péptica que han seguido un tratamiento, de que la sintomatología ulcerosa reaparezca en función del tiempo transcurrido desde la respuesta al tratamiento. Para ello, se somete al tratamiento a un conjunto de pacientes con úlcera péptica, siendo todos ellos fumadores. Al cabo de los 16 meses desde el primer caso de desaparición de la sintomatología se cierra el estudio, comprobando en cada paciente si la sintomatología había reaparecido o no en la última revisión y el tiempo transcurrido desde la respuesta al tratamiento y dicha revisión La fecha de inicio del tratamiento ha sido similar para todos los pacientes y, en consecuencia, al cierre del estudio en todos los casos ha habido respuesta al tratamiento. El máximo tiempo de respuesta fue de ocho semanas. Antes de comenzar el tratamiento algunos de los pacientes han decidido abandonar el hábito de fumar, por lo que se sospecha que, para un mismo tiempo transcurrido desde la respuesta al tratamiento, la probabilidad de que la sintomatología ulcerosa reaparezca depende del efecto del abandono del tabaco. Por otro lado, se sospecha que dicha probabilidad también depende de cuál ha sido el tiempo de respuesta al tratamiento Entonces, se desea estimar la probabilidad de reaparición de los síntomas en función del tiempo transcurrido desde la respuesta al tratamiento, conocido el tiempo de respuesta al tratamiento y el abandono o no del hábito de fumar. En este ejemplo, el conjunto de individuos son los pacientes en el hospital con úlcera péptica, la variable de interés cuyos valores son registrados a través del tiempo coincide con el evento de interés que es la reaparición de la sintomatología ulcerosa desde la respuesta al tratamiento y las variables predictoras son tiempo de tratamiento y hábito de fumar.

Otro tipo de estudios médicos en los que este tipo de análisis es utilizado es en aquellos en donde el interés radica en estimar la probabilidad de fallecimiento o muerte. Por ejemplo suponga que en un hospital a N pacientes desahuciados enfermos de leucemia se les aplica un tratamiento durante un periodo determinado esperando alargar el tiempo para la aparición del deceso. Entonces se desea estimar la probabilidad de que fallezcan en función del tiempo transcurrido desde la aplicación del tratamiento.

Por otro lado, supongamos que para los estudiantes de Ingeniería Química de una universidad Pública, se desea estimar la probabilidad de deserción en función del tiempo transcurrido desde su ingreso t_0. Para ello, se tiene información del conjunto de estudiantes de Ingeniería Química cuyo ingreso fue en t_0 (fecha similar para todos los estudiantes) y la variable de interés cuyos valores son registrados a través del tiempo coincide con el evento de interés que es la deserción. Del mismo modo podemos suponer que ahora el evento de interés es la no-retención, es decir, el no permanecer en la universidad en cada periodo de tiempo. Entonces, el interés radicaría en estimar, en función del tiempo, la probabilidad de que los estudiantes de Ingeniería Química de dicha universidad pública sean no - retenidos en la universidad.

En los tres ejemplos antes mencionados, el interés radica en estimar, en función del tiempo, la probabilidad de que un determinado evento de interés ocurra, al tiempo de ocurrencia se le conoce como *tiempo de sobrevivencia* y este evento puede ser la presencia de una enfermedad, la muerte o la deserción.

En su origen, los estudios de datos de sobrevivencia se concentraban en predecir la probabilidad de respuesta, sobrevivencia o tiempo promedio de vida, y en comparar las distribuciones de sobrevivencia de los resultados obtenidos de experimentos con animales o con pacientes humanos En los últimos años, la identificación del riesgo y/o factores de pronóstico relacionados con el evento de interés se han convertido igualmente importantes.

Algunos investigadores consideran el análisis de sobrevivencia como la aplicación de dos métodos estadísticos en un determinado problema. paramétrico si la distribución de los tiempos de sobrevivencia sigue una distribución de probabilidad conocida y no paramétrico si la distribución de los tiempos de sobrevivencia es desconocida. En el caso de los métodos paramétricos es necesario conocer exactamente los tiempos de sobrevivencia de cada uno de los sujetos de estudio y poder suponer que estos se pueden modelar con una distribución de probabilidad conocida tal como exponencial, weibull, lognormal, gamma, entre otras. Sin embargo, en la práctica generalmente no sucede por lo que hay que recurrir a otros métodos estadísticos como los no paramétricos.

2.4.1.2. Observaciones censuradas

Una de las principales ventajas del análisis de sobrevivencia es que considera los casos en que el evento de interés no ha ocurrido al final del estudio o en un tiempo de análisis determinado. El tiempo exacto de sobrevivencia de estos sujetos se desconoce. Estas reciben el nombre de *observaciones censuradas* o *tiempos censurados*. También puede ocurrir este tipo de observación

cuando no es posible por razones desconocidas seguir monitoreando a los individuos sujetos de estudio

Existen tres tipos de observaciones *censuradas*

1. Tipo I.

Cuando se observa el evento de interés por un determinado periodo definido de tiempo y aun no ha ocurrido y no es posible observarlo durante más tiempo se conoce como observaciones censuradas aquellas en las que al tiempo definido no ha ocurrido el evento observado También aquellas observaciones que no fue posible darles seguimiento por muerte o pérdida por otra razón distinta a la estudiado dentro del periodo definido.

2. Tipo II.

Se presentan en aquellas investigaciones en las cuales se fija una proporción meta de aparición del evento después de la cual se cierra el estudio. Las observaciones a las cuales no les sucedió el evento son las censuradas. Como en el caso anterior también son censuradas aquellas observaciones a las cuales por alguna razón no se les pudo dar seguimiento (se desconoce la información de ellas)

3. Tipo III

Es en aquellos estudios en los que el periodo esta definido y los sujetos de estudio entran a la investigación en distintos tiempos durante el periodo. Se les llama observaciones censuradas a las que al final del periodo de tiempo no les ha ocurrido el evento de interés y a aquéllas a las que no fue posible darles seguimiento.

Los dos primeros tipos de observaciones censuradas se les conoce como *Observaciones Censuradas Simples* y al tipo III como *Observaciones Censuradas Progresivas.* A este último tipo también se le conoce como *Observaciones Censuradas Aleatorias.* Los tres tipos de observaciones censuradas son *Censuradas a la Derecha.* Cuando no hay observaciones censuradas se dice que el conjunto de tiempos de sobrevivencia está completo.

Como se mencionó y justificó en el apartado 2.2 de este capítulo, el problema de la *deserción estudiantil* será abordado por un término alterno: *retención estudiantil.* Entonces, el objetivo de este trabajo es estimar, en función del tiempo, la probabilidad de que ocurra el evento crítico no-retención. El análisis de sobrevivencia en este contexto se denominaría análisis de retención ya que en el análisis de sobrevivencia el objetivo es estimar la probabilidad de muerte o presencia de determinada enfermedad en función del tiempo y en este trabajo en particular la muerte o presencia de determinada enfermedad equivale a la no - retención. Los datos censurados

(tipo 2) serán aquellos alumnos que al trimestre 12 aún permanecen en la universidad y por tanto el evento crítico no ha sido observado. Las observaciones censuradas son alumnos con carácter de activo rezagado.

2.4.1.3. Funciones de Tiempo de Sobrevivencia.

Los tiempos de sobrevivencia miden el tiempo a cierto evento tal como respuesta a un tratamiento, muerte, el desarrollo de una enfermedad dada, divorcio entre otros. La distribución de los tiempos de sobrevivencia es generalmente descrito o caracterizado con tres funciones:

1. La función de sobrevivencia.
2. La función de densidad de probabilidad.
3. La función de riesgo (hazard function).

En la práctica, las tres funciones de tiempo de sobrevivencia pueden ser usadas para ilustrar distintos aspectos de los datos. El problema central en el análisis de sobrevivencia es la estimación en los datos muestrales de una o más de estas tres funciones de sobrevivencia y realizar inferencias acerca del comportamiento de la sobrevivencia en la población.

Sea T el tiempo de sobrevivencia. La distribución de T puede ser caracterizada, como ya se mencionó, por las tres siguientes funciones equivalentes .

1. Función de sobrevivencia

Esta función, denotada por S(t), se define como la probabilidad de que un individuo sobreviva más allá del tiempo t.

$$S(t) = P(\text{un individuo sobreviva más allá del tiempo } t)$$

$$S(t) = P(T > t)$$

De la definición de la función de distribución F(t) de T,

$$S(t) = 1 - P(\text{un individuo no sobreviva antes del tiempo } t)$$

$$S(t) = 1 - P(T \leq t) \qquad (1)$$

$$S(t) = 1 - F(t) \qquad (2)$$

Por tanto la función de sobrevivencia $S(t)$ es el complemento de la función de distribución acumulada. Es decir, $S(t)$ es una función de tiempo t no creciente y con las propiedades de,

$S(t) = 1$ para $t = 0$

$S(t) = 0$ para $t = \infty$

Esto es, la probabilidad de sobrevivencia en al menos el tiempo 0 es 1, es decir, al inicio del estudio el evento de interés no ha ocurrido en ninguna de las observaciones. Conforme el tiempo transcurre la probabilidad de sobrevivencia decrece hasta que para t suficientemente grande, la función de sobrevivencia es 0.

Por otro lado, si en t=0 la probabilidad de sobrevivencia es 1 entonces en t=0 la función de distribución acumulada es 0 dado que al inicio del estudio la probabilidad de no sobrevivir es 0. En el mismo sentido, si en $t = \infty$ la probabilidad de sobrevivencia es 0 entonces la función de distribución acumulada es 1 ya que la probabilidad de no sobrevivir al evento de interés para t grande es 1.

La función S(t) es también conocida como la *tasa acumulada de sobrevivencia*. La representación gráfica de la función de sobrevivencia es la *curva de sobrevivencia*.

La función de sobrevivencia es utilizada para encontrar el percentil 50 (mediana) y otros percentiles tales como el 25 o el 75 del tiempo de sobrevivencia y para comparar las distribuciones de sobrevivencia de dos o más grupos. También puede ser calculado el valor medio sin embargo, éste sólo es usado como medida de tendencia central de la distribución ya que un pequeño número de individuos con excepcionalmente cortos o largos periodos de vida pueden causar que el tiempo promedio de sobrevivencia sea grande o pequeño.

En la práctica si no hay observaciones censuradas, la función de sobrevivencia puede ser estimada como la proporción de pacientes que sobreviven más allá del tiempo t:

$$\hat{S}(t) = \frac{\text{Número de pacientes que sobreviven más allá del tiempo t}}{\text{Número total de pacientes}} \qquad (3)$$

donde $\hat{S}(t)$ es la función de sobrevivencia estimada.

Cuando hay observaciones censurados el numerador de la ecuación anterior no puede ser siempre calculado. En estos casos, se recurre a las estimaciones no paramétricas de $\hat{S}(t)$

2. Función de densidad de probabilidad.

Como en el casos de cualquier otra variable aleatoria continua, el tiempo de sobrevivencia T tiene una función de densidad de probabilidad definida como el límite de la probabilidad de que un individuo fallezca en el intervalo corto de tiempo (t, t+Δt) o simplemente la probabilidad de no sobrevivir en un pequeño intervalo por unidad de tiempo

$$f(t) = \lim_{\Delta t \to 0} \frac{P(\text{Un individuo fallezca en el intervalo } (t, t + \Delta t))}{\Delta t} \qquad (4)$$

La gráfica de f(t) es conocida como curva de densidad La función de densidad tiene las siguientes dos propiedades:

1. f(t) es una función no negativa:

$f(t) \geq 0$ para toda $t \geq 0$

$f(t) = 0$ para $t < 0$

2. El área entre la curva de densidad y el eje x es 1.

En la práctica si no hay observaciones censuradas, la función de probabilidad de densidad f(t) es estimada como la proporción de pacientes falleciendo en un intervalo por unidad de longitud de tiempo.

$$\hat{f}(t) = \frac{\text{Número de paciente que fallecen en el intervalo que empieza al tiempo } t}{(\text{Número total de pacientes})(\text{Longitud del intervalo})} \qquad (5)$$

Como en el caso de la función de sobrevivencia, la función de densidad no puede ser calculada de esta manera si existen casos censurados.

3. Función de riesgo

La *función de riesgo* h(t) del tiempo de sobrevivencia T proporciona la tasa condicional de fallecimiento. Esto es definido como la probabilidad de fallecimiento al comienzo del intervalo, o como el límite de la probabilidad de que un individuo fallezca en un pequeño intervalo de tiempo, (t,t+Δt) dado que el individuo ha sobrevivido al tiempo t.

$$h(t) = l \lim_{\Delta t \to 0} \frac{P(\text{un individuo que no presentó el evento en tiempo } T \leq t \text{ fallezca en el intervalo de tiempo } (t, t + \Delta t))}{\Delta t} \tag{6}$$

La *función de riesgo* puede ser también definida en términos de la función de distribución F(t) y la función de densidad f(t).

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{7}$$

Si recordamos la definición (2) de función de sobrevivencia, entonces la expresión anterior se puede escribir como $h(t) = \frac{f(t)}{S(t)}$, es decir, la función hazard es el cociente entre la función de densidad de probabilidad y la función de sobrevivencia.

La *función de riesgo* es también conocida como la tasa instantánea de muerte, fuerza de mortalidad, tasa de mortalidad condicional y tasa de muerte en edad especifica Es una medida de propensión a la muerte en el sentido que la cantidad Δt h(t) es la proporción esperada de individuos a la edad t que fallecerán en el corto periodo de tiempo (t,t+Δt). La *función de riesgo* proporciona el riesgo de fallecer por unidad de tiempo durante el proceso de edad

En la práctica cuando no hay observaciones censuradas, la *función de riesgo* es estimada como la proporción de pacientes muriendo en un intervalo por unidad de tiempo dado que han sobrevivido al principio del intervalo.

$$\hat{h}(t) = \frac{\text{Numero de pacientes muriendo en el intervalo que empieza en el tiempo t}}{(\text{Número de pacientes que sobreviven al tiempo t})(\text{Longitud del intervalo})}$$

$$\hat{h}(t) = \frac{\text{Numero de pacientes muriendo por unidad de tiempo en el intervalo}}{(\text{Número de pacientes que sobreviven al tiempo t})} \tag{8}$$

En Ciencias Actuariales frecuentemente se utiliza la *función de riesgo* promedio en el intervalo en el cual el número de pacientes muriendo por unidad de tiempo en el intervalo es dividido por el número promedio de sobrevivientes en el punto medio del intervalo.

$$\hat{h}(t) = \frac{\text{Numero de pacientes muriendo por unidad de tiempo en el intervalo}}{\text{(Número de pacientes que sobreviven al tiempo t) - .5 (Número de muertes en el intervalo)}} \qquad (9)$$

La estimación actuarial anterior proporciona una estimación de la *función de riesgo* más alta y por tanto una estimación más conservadora.

La *función de riesgo* puede crecer, decrecer permanecer constante o puede indicar un proceso combinado.

La *función de riesgo* acumulada es definida como:

$$H(t) = \int_0^t h(x)\,dx \qquad (10)$$

Entonces,

si t= 0, S(t)=1 luego entonces H(t)=0,

si t=∞, S(t)=0 luego entonces H(t)=∞

- Relación entre las funciones de sobrevivencia.

Las tres funciones definidas anteriormente son matemáticamente equivalentes. Dada una de ellas, las otras dos pueden ser derivadas.

1. Sustituyendo (2) en (7) se tiene que,

$$h(t) = \frac{f(t)}{S(t)} \qquad (11)$$

Debido a que la función de densidad es la derivada de la función de distribución,

F(t) = 1 – S(t)

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t) \qquad (12)$$

2. Sustituyendo la expresión anterior en la ecuación (11) se tiene,

$$h(t) = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\ln S(t) \qquad (13)$$

Integrando de 0 a t y utilizando que S(0)=1,

$$-\int_0^t h(x)\,dx = \ln(S(t))$$

o,

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

o,

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0 h(x)dx\right] \qquad (14)$$

De la expresión (11) y (14) se obtiene,

$$f(t) = h(t)\exp[-H(t)] \qquad (15)$$

Por tanto, si f(t) es conocida, la función de sobrevivencia S(t) puede ser obtenida por la relación básica entre f(t), F(t) y (2). La *función de riesgo* puede ser determinada por (12) y (13).

Si S(t) es conocida, f(t) y h(t) pueden ser determinadas por (12) y (11) respectivamente o primero derivar h(t) de (13) y después f(t) de (11). Si h(t) es dada, S(t) y f(t) pueden ser obtenidas respectivamente de (14) y (15).

Por tanto, dada cualquiera de las tres funciones de sobrevivencia, las otras dos pueden ser fácilmente derivadas.

2.4.1.4. Método no paramétrico producto - límite de Kaplan y Meier

Los métodos no paramétricos son menos eficientes que los métodos paramétricos cuando los tiempos de sobrevivencia siguen una distribución de probabilidad teórica conocida pero son más eficientes cuando se desconoce dicha distribución. En la práctica, los métodos no paramétricos son más utilizados porque su aplicación e interpretación es más sencilla además de que el uso de métodos paramétricos es limitado dado que en general se desconoce la distribución de probabilidad de los tiempos de sobrevivencia.

De las tres funciones de tiempo de sobrevivencia descritas en el apartado anterior 2 4 1.3 , la función de sobrevivencia o su representación gráfica, curva de sobrevivencia, es la más utilizada. Esta función puede estimarse por el método no paramétrico producto - límite desarrollado por Kaplan y Meier (1958). Este método tiene una variante que es la estimación de las tablas de vida, la única diferencia entre esta y el método producto - límite de Kaplan y Meier es que en el primer caso los tiempo de sobrevivencia están agrupados en intervalos. El método no paramétrico producto - límite de Kaplan y Meier puede ser considerado como un caso especial de la estimación de las tablas de vida donde cada intervalo tiene una sola observación. A continuación se describe la estimación producto – límite de Kaplan y Meier

Consideremos primero la situación en la que en todos los casos se observa el evento crítico y por tanto se conocen los tiempos exactos de sobrevivencia. Sea $t_1, t_2, ..., t_n$ los tiempos exactos de sobrevivencia de los n individuos en estudio. Conceptualmente, se considera este grupo de individuos como muestra aleatoria de una población más grande de individuos similares Se renombran los n tiempos de sobrevivencia $t_1, t_2, ..., t_n$ en orden ascendente tal que,

$$t_{(1)} \leq t_{(2)} \leq ... \leq t_{(n)}$$

Según la definición (1) y (2), la función de sobrevivencia en $t_{(i)}$ puede ser estimada como,

$$\hat{S}(t_{(i)}) = \frac{n-i}{n} = 1 - \frac{i}{n} \qquad (16)$$

donde $n-i$ es el número de individuos en la muestra que sobreviven más allá del tiempo $t_{(i)}$. Si dos o más $t_{(i)}$ son iguales (observaciones empatadas), el valor más grande de i es utilizado. Por ejemplo, si $t_{(2)} = t_{(3)} = t_{(4)}$, entonces,

$$\hat{S}(t_{(2)}) = \hat{S}(t_{(3)}) = \hat{S}(t_{(4)}) = \frac{n-4}{n}$$

Lo anterior proporciona una estimación conservadora de las observaciones empatadas.

Debido a que todos los individuos están vivos o no han experimentado el evento crítico al inicio del estudio y ninguno sobrevive más allá que $t_{(n)}$ se tiene que,

$$\hat{S}(t_{(a)}) = 1 \qquad \text{y} \qquad \hat{S}(t_{(n)}) = 0 \tag{17}$$

En la práctica, $\hat{S}(t)$ se calcula en todos los diferentes tiempo de sobrevivencia De este modo se abarca los casos de los intervalos entre los distintos tiempos de sobrevivencia en los cuales ningún individuo experimenta el evento crítico y $\hat{S}(t)$ permanece constante Las ecuaciones (16) y (17) muestran que $\hat{S}(t)$ es una función de salto que empieza en 1 y que decrece en saltos de $\frac{1}{n}$ (si no hay empates) hasta cero. Cuando $\hat{S}(t)$ es graficada con t, los percentiles del tiempo de sobrevivencia pueden ser leídos de la gráfica o calculados de $\hat{S}(t)$.

Teóricamente, $\hat{S}(t)$ puede ser graficada como una función de salto dado que permanece constante entre dos tiempos observados de sobrevivencia.

Este método sólo puede ser utilizado si todos los individuos experimentan el evento crítico Si alguno de ellos no ha experimentado el evento crítico al final del estudio entonces el método de producto - límite de Kaplan y Meier es utilizado para la estimación de $\hat{S}(t)$. La base de este método será ilustrada en el siguiente ejemplo 1.

Ejemplo 1.

Suponga que 10 pacientes ingresan a una clínica que realiza un estudio al principio de 1988; durante el año seis mueren y cuatro sobreviven Al final del año otros veinte pacientes se añaden al estudio. En 1989, tres pacientes que entraron al inicio de 1988 y doce que entraron después fallecen, dejando uno y cinco sobrevivientes respectivamente. Suponga que el estudio termina al final de 1989 y que se desea estimar la proporción de pacientes en la población que sobrevivirán por dos años o más, es decir, S(2).

El primer grupo de pacientes en este ejemplo fue monitoreado por dos años mientras que el segundo grupo sólo por 1 año Una posible estimación, *la estimación reducida de la muestra*, es $\hat{S}(2) = \frac{1}{10} = .1$, la cual ignora los 20 pacientes que solo estuvieron en el estudio por un año Kaplan y Meier afirman que la segunda muestra, que sólo estuvo en observación por un año, puede contribuir a la estimación de S(2).

Pacientes que sobrevivieron dos años deben ser considerados como sobrevivientes en el primer año y sobrevivientes en más de un año Entones, la probabilidad de sobrevivir 2 años o más es igual a la probabilidad de sobrevivir el primer año y sobrevivir más de un año Esto es,

$S(2) =$ P(sobrevivir el primer año y después sobrevivir más de un año)

que puede escribirse,

$S(2) =$ P(sobrevivir dos años dado que el paciente ha sobrevivido el primer año)
X P(sobrevivir el primer año)

(18)

La estimación producto – límite de Kaplan y Meier de S(2) siguiendo (18),

$\hat{S}(2) =$ (proporción de pacientes que sobrevivieron dos años dado que sobrevivieron 1 año) **X** (proporción de pacientes que sobrevivieron 1 año)

(19)

En el ejemplo, uno de los cuatro pacientes que sobrevivió el primer año, sobrevivió dos años entonces la proporción del primer factor de la expresión (19) es $\frac{1}{4}$. Cuatro de los diez pacientes que entraron al principio de 1988 y cinco de los 20 pacientes que entraron al final de 1988 sobrevivieron un año. Entonces, el segundo factor de la expresión (19) es $\frac{(4+5)}{(10+20)}$. La estimación producto – límite de Kaplan y Meier de S(2) es,

$$\hat{S}(2) = \frac{1}{4} * \frac{4+5}{10+20} = 0.25 * .3 = .075$$

La regla se generaliza del siguiente modo. La probabilidad de sobrevivir k (≥ 2)o más años desde el inicio del estudio es el producto de las k tasas observadas de sobrevivencia

$$\hat{S}(k) = p_1 * p_2 * p_3 * ... * p_k \qquad (20)$$

donde p_1 denota la proporción de pacientes que sobreviven al menos 1 año

p_2 denota la proporción de pacientes que sobreviven el segundo año, después de que sobrevivieron un año.

p_3 denota la proporción de pacientes que sobreviven el tercer año, después de sobrevivieron dos años.

p_k denota la proporción de pacientes que sobreviven el k-ésimo año ya que han sobrevivido k-1 años.

Entonces la estimación producto – límite Kaplan y Meier de la probabilidad de sobrevivir cualquier número de particular de años desde el inicio del estudio es el producto de la estimación en ese año por la estimación del año anterior. La tasa observada de sobrevivencia para un año en particular es,

$$\hat{S}(t) = \hat{S}(t-1)p_t \qquad (21)$$

Los estimadores producto – límite de Kaplan y Meier son los estimadores de máxima verosimilitud.

En la práctica la estimación producto – límite de Kaplan y Meier puede ser calculada construyendo una tabla con las siguientes cinco columnas:

1. La primera columna t, debe contener los tiempos de sobrevivencia, tanto censurados como no censurados, ordenados del menor al mayor. A las observaciones censuradas se les añade un signo + para diferenciarlas de las no censuradas Si una observación censurada tiene el mismo valor que una no censurada, la no censurada debe aparecer primero.

2 La segunda columna i, presenta los rangos correspondientes a las observaciones de la columna 1.

3. La tercera columna r, son los rangos de las observaciones no censuradas

4. Calcular $\frac{(n-r)}{(n-r+1)}$ o p_i para cada observación no censurada $t_{(i)}$ que resulta la proporción de casos que sobreviven más allá de $t_{(i)}$.

5 En la última columna, $\hat{S}(t)$ es el producto de todos los valores de $\frac{(n-r)}{(n-r+1)}$ hasta t Si algunas observaciones no censuradas son empates, el valor más pequeño de $\hat{S}(t)$ es utilizado

En resumen, este proceso puede describirse como:

- Sea n el número total de individuos para los cuales sus tiempos de sobrevivencia, ya sean censurados o no censurados, están disponibles

- Renombre los n tiempo de sobrevivencia en orden creciente tal que $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(n)}$ Entonces,

$$\hat{S}(t) = \prod_{t_{(r)} \leq t} \frac{n-r}{n-r+1} \qquad (22)$$

donde r corre en los enteros positivos para los cuales $t_{(r)} \leq t$ y $t_{(r)}$ es no censurada Los valores de 2 son enteros consecutivos 1,2,...,n si no hay observaciones censuradas La mediana estimada del tiempo de sobrevivencia es el percentil 50, el cual es el valor de t cuando $\hat{S}(t) = 0.50$.

A continuación se presenta el ejemplo 2 que ilustra los cálculos anteriores.

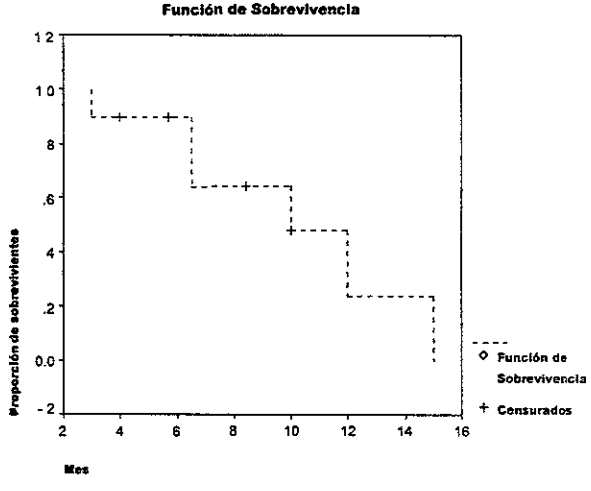
Suponga que un grupo de 10 pacientes desahuciados con cáncer terminal se somete a un nuevo tratamiento, 6 de ellos mueren en los tiempos 3,6 5,6.5,10, 12 y 15 meses, a uno de los pacientes ya no fue posible seguir observándolo después del tiempo 8 4 meses y 3 pacientes están aún vivos al final de estudio pero cada uno se sometió al tratamiento por diferentes periodos de tiempo. 4.5.7 y 10 meses.

La tabla 1 muestra el calculo de $\hat{S}(t)$ y la gráfica 1 ilustra la curva de sobrevivencia $\hat{S}(t)$.

Tabla 1. Cálculo del producto-límite de Kaplan y Meier de $\hat{S}(t)$

Recaída T	Rango I	R	$\frac{(n-r)}{(n-r+1)}$	$\hat{S}(t)$
3.0	1	1	$\frac{9}{10}$	$\frac{9}{10} = .9$
4 0+	2	-	-	-
5.7+	3	-	-	-
6 5	4	4	$\frac{6}{7}$	$\frac{9}{10} * \frac{6}{7} = 0\ 771$
6.5	5	5	$\frac{5}{6}$	$\frac{9}{10} * \frac{6}{7} * \frac{5}{6} = 0.643$
8.4+	6	-	-	-
10.0	7	7	$\frac{3}{4}$	$\frac{9}{10} * \frac{6}{7} * \frac{5}{6} * \frac{3}{4} = 0.482$
10.0+	8	-	-	-
12.0	9	9	$\frac{1}{2}$	$\frac{9}{10} * \frac{6}{7} * \frac{5}{6} * \frac{3}{4} * \frac{1}{2} = 0.241$
15.0	10	10	0	0

Gráfica 1. Función $\hat{S}(t)$ del ejemplo 2.



La mediana estimada es m = 9 8 meses

La estimación de $\hat{S}(t)$ en t = $t_{(i)}$ está relacionado con $\hat{S}(t)$ en t = $t_{(i-1)}$ y puede ser escrito como:

$$\hat{S}(t_{(i-1)}) = S(t_{(i-1)}) \frac{n-t}{n-t+1}$$

La varianza de la estimación del producto - límite de la estimación $\hat{S}(t)$ es aproximadamente:

$$Var[\hat{S}(t)] = [\hat{S}(t)]^2 \sum_r \frac{1}{(n-r)(n-r+1)} \qquad (23)$$

donde r incluye los enteros positivos tales que $t_{(r)} \leq t$ y $t_{(r)}$ corresponde al fallecimiento

Para los datos del ejemplo,

$$Var[\hat{S}(10)] = [0.482]^2 \left[\frac{1}{9*10} + \frac{1}{6*7} + \frac{1}{5*6} + \frac{1}{3*4} \right] = 0.0352$$

y el error estándar estimado es 0.1876.

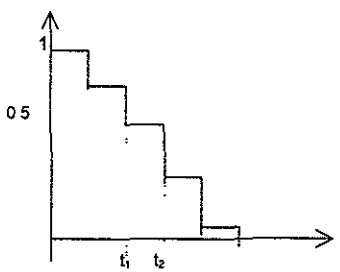
El método de Kaplan y Meier proporciona estimaciones muy útiles de la probabilidad de sobrevivencia y la representación gráfica de la distribución de la curva de sobrevivencia Es el método más comúnmente utilizado en el análisis de sobrevivencia Sin embargo, este método tiene algunas particularidades que es necesario enunciar:

1. La estimación Kaplan y Meier es limitada al intervalo de tiempo en el que las observaciones fallecen. Si la observación con tiempo de sobrevivencia mayor es no censurada, el método producto-límite de Kaplan y Meier es cero en ese tiempo. Esta estimación es correcta dado que nadie en la muestra vive más tiempo. Si el tiempo de observación más grande es censurado, la estimación del método producto-límite de Kaplan y Meier no puede ser nunca cero y es indefinido más allá de la observación con tiempo de sobrevivencia más grande.

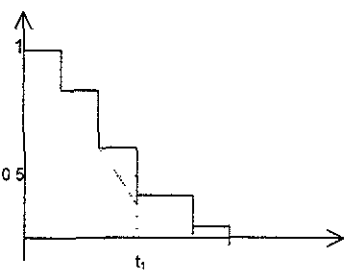
2. La medida más utilizada en el análisis de sobrevivencia es la mediana del tiempo de sobrevivencia. Una simple estimación de la mediana puede ser leída de las curvas de sobrevivencia estimadas por el método producto-límite de Kaplan y Meier como el tiempo t en

el que $\hat{S}(t) = 0.5$. Sin embargo la solución puede ser no única Considere la siguiente gráfica 2 donde la curva de sobrevivencia es horizontal en $\hat{S}(t) = 0.5$, cualquier valor t en el intervalo t_1 a t_2 es una estimación razonable de la mediana Una solución practica es tomar el punto medio del intervalo como la estimación producto-límite de la mediana La gráfica 3 presenta un caso diferente en el cual la estimación t_1 tiende a sobrestimar la mediana Una manera práctica de manejar este problema es conectando los puntos y después calcular la mediana.

Gráfica 2. Estimación de la mediana de sobrevivencia con el método producto-límite. de Kaplan y Meier



Gráfica 3. Estimación de la mediana de sobrevivencia con el método producto-límite de Kaplan y Meier.



3. Si menos del 50% de las observaciones son no censuradas y el tiempo de sobrevivencia más grande es censurado, la mediana del tiempo de sobrevivencia no puede ser estimada Una manera práctica de manejar esta situación es usando la probabilidad de sobrevivir en un determinado periodo de tiempo, digamos uno , tres o cinco años o el tiempo promedio de sobrevivencia limitado a un dado tiempo t.

4. El método producto-límite de Kaplan y Meier supone que el ser observación censurada es independiente de los tiempos de sobrevivencia, En otras palabras, la razón por la cual una observación es censurada no está relacionada a la causa de muerte. Este supuesto es cierto si el paciente está aún vivo al final del periodo de estudio. Sin embargo, este supuesto es violado si el paciente desarrolló severos efectos adversos causados por el tratamiento y es forzado a abandonar el estudio antes de la muerte. Cuando hay una inapropiada manera de establecer que una observación es censurada, el método producto-límite de Kaplan y Meier no es apropiado.

5. Tal como para otros estimadores, el error estándar (SE) del estimador Kaplan y Meier de $S(t)$ proporciona un indicador del error potencial de $\hat{S}(t)$. Suponiendo normalidad de $\hat{S}(t)$, el intervalo de confianza al 95% para $S(t)$ es $\hat{S}(t) \pm 1\,96$ (SE) $\hat{S}(t)$

2.4.1.5. Métodos no paramétricos para la comparación de distribuciones de sobrevivencia

Como ya se menciono en el apartado 2.4.1.4 el método no paramétrico producto - límite de Kaplan y Meier estima la función de sobrevivencia, es decir, la probabilidad de que ocurra determinado evento crítico en función del tiempo. El interés radica, entonces, en conocer si existen diferencias significativas entre las curvas de sobrevivencia de diferentes grupos Por ejemplo, determinar si hay diferencias significativas entre la probabilidad de que el evento crítico muerte ocurra para el grupo de estudiantes mujeres y para el grupo de estudiantes hombres Para esto, existen diferentes pruebas no paramétricas tales como la prueba Gehan, Log Rank y Mantel Haenzel para la comparación de funciones de sobrevivencia entre dos grupos y la prueba Breslow, Tarone Ware y Log Rank para el caso general de k grupos.

En este apartado 2.4.1.5. se desarrollará el caso más simple en el que se compara la función de sobrevivencia de dos grupos.

Suponga que hay n_1 y n_2 individuos que reciben los tratamientos 1 y 2 respectivamente. Sea $x_1,...,x_{r_1}$ las r_1 observaciones que fallecen y $x^+_{r_1+1},...,x^+_{n_1}$ los $n_1 - r_1$ observaciones censuradas en el grupo 1. En el grupo 2, sea $y_1,...,y_{r_2}$ las r_2 observaciones que fallecen y $y^+_{r_2+1},...,y^+_{n_2}$ los $n_2 - r_2$ observaciones censuradas. Esto es, al final del estudio, $n_1 - r_1$ individuos que recibieron el tratamiento 1 y $n_2 - r_2$ individuos que recibieron el tratamiento 2 aún están vivos

Suponga que las observaciones en el grupo 1 son muestras de una distribución con función de sobrevivencia $S_1(t)$ y las observaciones del grupo 2 son muestras de una distribución con función de sobrevivencia $S_2(t)$. Entonces la hipótesis nula a considerar es,

$H_0 : S_1(t) = S_2(t)$ (tratamientos 1 y 2 son igualmente efectivos)

contra la hipótesis alterna

$H_1 : S_1(t) > S_2(t)$ (tratamiento 1 es más efectivo que el tratamiento 2)

o

$H_2 : S_1(t) < S_2(t)$ (tratamiento 2 es más efectivo que el tratamiento 1)

o

$H_3 : S_1(t) \neq S_2(t)$ (tratamiento 1 y 2 no son igualmente efectivos)

Cuando no hay observaciones censuradas, pruebas no paramétricas tales como la prueba de Wilcoxon, la prueba de U-Mann y Whitney y la prueba de los signos pueden ser usadas para comparar las dos distribuciones de sobrevivencia.

1. Prueba Gehan (generalización de la prueba Wilcoxon)

En esta prueba todas las observaciones x_i o x_i^+ en el grupo 1 es comparado con cada una de las observaciones y_j o y_j^+ en el grupo 2 y un puntaje U_{ij} es asignado al resultado de cada una de las comparaciones. Para ilustrar, supongamos que la hipótesis alternativa es $H_1 : S_1(t) > S_2(t)$, es decir. El tratamiento 1 es más efectivo que el tratamiento 2

Sea

$$U_{ij} = \begin{cases} +1 & \text{si } x_i > y_j, x_i \geq y_j^+ \\ 0 & \text{si } x_i = y_j \text{ o } x_i^+ < y_j \text{ o } y_j^+ < x_i \text{ o } (x_i^+, y_j^+) \\ -1 & \text{si } x_i < y_j \text{ o } x_i \leq y_j^+ \end{cases}$$

y calcular la estadística de

$$w = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} U_{ij} \qquad (24)$$

donde la suma es sobre todas las comparaciones n_1 n_2 Entonces, hay una contribución de la estadística W para todas las comparaciones donde ambas observaciones fallecen (excepto para empates) y para cada una de las comparaciones en donde las observaciones censuradas son igual o mayor que un fallecimiento.

El calculo de W es laborioso cuando n_1 y n_2 son grandes Mantel(1967) muestra que W puede ser calculado de forma alternativa asignando un puntaje para cada observación basándose en su rango relativo. En la prueba de Gehan cada observación de la muestra 1 es comparada con cada observación de la muestra 2 En el procedimiento de Mantel, las 2 muestras se combinan en una única muestra de $n_1 + n_2$ observaciones y se compara cada observación con las restantes $n_1 + n_2 - 1$

Entonces, sea U_i con $i = 1,...,n_1 + n_2$. La U_i de $n_1 + n_2$ está definida para una población finita con media cero y entonces con el procedimiento de Mantel, el estadístico de Gehan es,

$$W = \sum_{i=1}^{n_1} U_i \tag{25}$$

donde la suma es sobre todas las U_i de la muestra 1 únicamente. Ya sea de (24) o (25) , *es claro que W es un número grade positivo si* H_1 es cierta. Mantel también sugiere que la varianza permutacional de W puede ser usada, en vez de la más complicada formula de varianza deducida por Gehan. La distribución permutacional de W puede ser obtenida considerando todas las manera de seleccionar aleatoriamente n_1 de U_1,

$$\binom{n_1 + n_2}{n_2} = \frac{(n_1 + n_2)!}{n_1! n_2!} \tag{26}$$

La prueba estadística W bajo H_0 puede ser considerada aproximadamente distribuida normalmente con media cero y varianza

$$\text{var}(W) = \frac{n_1 n_2 \sum_{i=1}^{n_1+n_2} U_i^2}{(n_1 + n_2)(n_1 + n_2 - 1)} \tag{27}$$

Debido a que W es discreta, una corrección apropiada de continuidad de 1 es generalmente usada cuando no hay empates ni observaciones censuradas. De otro modo, una corrección de continuidad de .5 podría ser apropiada

Dado que W tienen una distribución normal asintótica con media cero y varianza en (27),

$Z = \frac{W}{\sqrt{\text{var}(W)}}$ tiene una distribución normal estándar. La zona de rechazo es $Z > Z_\alpha$ para H_1, $Z < -Z_\alpha$ para H_2, y $|Z| > Z_{\alpha/2}$ para H_3, donde $P(Z > Z_\alpha \mid H_O) = \alpha$

El número U_i puede ser calculado en dos etapas. La primera etapa consiste, para cada observación, la unidad más el número de observaciones restantes que es mayor que, es decir, R_{1i}. La segunda etapa consiste en R_{2i}, que es la unidad más el número de observaciones restantes y que la observación particular es definitivamente menor que. Sea $U_i = R_{1i} - R_{2i}$. El calculo de R_{1i} y R_{2i} puede ser realizado sistemáticamente en pasos como se ilustra en el siguiente ejemplo 3

Ejemplo 3:

Las mujeres pacientes con cáncer de pecho son seleccionadas aleatoriamente para recibir CMF (administración cíclica de ciclofosfamide, metatrexate y flouroracil) o ningún tratamiento después de una mastectomía. Al final de dos años, los tiempos de recaída en meses se registran del siguiente modo:

CMF (grupo 1)	23,16+,18+,20+,24+
Control (grupo 2)	15,18,19,19,20

La hipótesis nula y la alternativa es

$H_0 : S_1 = S_2$ (dos tratamientos igualmente efectivos)

$H_1 : S_1 > S_2$ (CMG más eficiente que no tratamiento)

El cálculo de R_{1i} y R_{2i} y U_i se presentan en la tabla 2. Entonces W=1+2+5+4+6=18, Var(W)=(5)(5)(208)/(10)(9) = 57.78, y $Z = 18/\sqrt{57.78} = 2.368$ Suponga que el nivel de significación usado es α =0.05, $Z_{0.05} = 1.64$. Entonces el valor calculado de Z está en la zona de rechazo Entonces, se rechaza la hipótesis nula al nivel de 0 05 y se concluye que los datos muestran que CMF es más efectivo que no recibir ningún tratamiento. De hecho, aproximadamente el valor p correspondiente a Z = 2.368 es 0 009.

La suma de U_i es igual a cero. Este hecho puede ser utilizado para verificar el procedimiento de calculo

Tabla 2. Procedimiento de Mantel para calcular U_i en la prueba de Gehan (generalización de Wilcoxon)

	Observaciones de las dos muestras en orden ascendente	15	16*	18	18*	19	19	20	20*	23	24*
	Calculo de R_{1i}										
Paso 1	Rango de la izquierda a la derecha omitiendo observaciones censuradas	1		2		3	4	5		6	
Paso 2	Asignación del siguiente rango más alto a observaciones censuradas		2		3				6		7
Paso 3	Reducción del rango de las observaciones empatadas al menor rango para el valor.						3				
Paso 4	R_{1i}	1	2	2	3	3	3	5	6	6	7
	Calculo de R_{2i}										
Paso 5	Rango de derecha a izquierda	10	9	8	7	6	5	4	3	2	1
Paso 6	Reducción del rango de las observaciones empatadas al menor rango para el calor					5					
Paso 7	Reducción del rango de las observaciones censuradas a 1		1		1				1		1
Paso 8	R_{2i}	10	1	8	1	5	5	4	1	2	1
	$U_i = R_{1i} - R_{2i}$	-9	1[a]	-6	2[a]	-2	-2	1	5[a]	4[a]	6[a]

[a] Para el grupo 1

La generalización de la prueba de Gehan se conoce como prueba Breslow. La hipótesis nula de la prueba Breslow es:

$$H_0 : S_1(t) = S_2(t) = \ldots = S_i(t)$$

2. Prueba Logrank.

Para introducir la prueba Log Rank primero se describirá la prueba Cox Mantel dado que en ella se ilustra algunos de las definiciones utilizadas en la prueba Log Rank.

- Prueba Cox-Mantel

Sea $t_{(1)} < \ldots < t_{(k)}$ los distintos tiempo de fallecimiento en los dos grupos juntos y $m_{(i)}$ el numero de tiempos de fallecimientos iguales a t_i, o la multiplicidad de t_i, tal que

$$\sum_{i=1}^{k} m_{(i)} = r_1 + r_2 \tag{28}$$

Además, sea $R(t)$ el conjunto de individuos que aún están expuestos al riesgo de fallecimiento al tiempo t y cuyos tiempos de fallecimiento o censura son al menos t. $R(t)$ se le denomina el conjunto en riesgo al tiempo t. Sea n_{1i} y n_{2i} el número de pacientes en $R(t)$ que pertenecen a los grupos de tratamientos 1 y 2 respectivamente. El número total de observaciones, fallecimientos u observaciones censuradas en $R(t_{(i)})$ es $r_{(i)} = n_{(i)} = n_{1i} + n_{2i}$.

Se define

$$U = r_2 - \sum_{i=1}^{k} m_{(i)} A_{(i)} \tag{29}$$

$$I = \sum_{i=1}^{k} \frac{m_{(i)}\left(r_{(i)} - m_{(i)}\right)}{r_{(i)} - 1} A_{(i)}(1 - A_{(i)}) \tag{30}$$

donde $r_{(i)}$ es el número de observaciones, fallecimientos u observaciones censuradas en $R(t_{(i)})$ y $A_{(i)}$ es la proporción de $r_{(i)}$ que pertenecen al grupo 2. Una prueba asintótica de dos muestras se obtiene entonces calculando el estadístico $C = \frac{U}{\sqrt{I}}$ como normal estándar bajo la hipótesis nula. El siguiente ejemplo 4 ilustra el procedimiento.

Ejemplo 4.

Considere los datos del ejemplo de la prueba Gehan Hay k = 5 distintos tiempos de fallecimientos en los dos grupos, $r_1 = 1$ y $r_2 = 5$. En la siguiente tabla se muestra la información necesaria para llevar a cabo la prueba Cox-Mantel

Tabla 3. Prueba Cox-Mantel

		Conjunto en riesgo al tiempo t			
		Muestra 1	Muestra 2		
Distintos tiempos de fallecimiento t_i	$m_{(i)}$	n_{1i}	n_{2i}	$r_{(i)}$	$\hat{A}_{(i)}$
15	1	5	5	10	0 5
18	1	4	4	8	0 5
19	2	3	3	6	0 5
20	1	3	1	4	0 25
23	1	2	0	2	0

$$U = 5 - (0.5 + 0.5 + 2 * 0.58 + 0.25) = 5 - 2.25 = 2.75$$

$$I = \frac{1*9}{9}(0.5*0.5) + \frac{1*7}{7}(0.5*0.5) + \frac{2*4}{5}(0.5*0.5) + \frac{1*3}{3}(0.25*0.75)$$

$$I = 0.25 + 0.25 + 00.4 + 0.1875 = 1.0875$$

Entonces, $C = \frac{2.75}{1.0875} 2.6 > Z_{0\,05} = 1.64$ y se rechaza la hipótesis nula H_0 al nivel 0.05. El p-valor correspondiente a $Z = 2.637$ es aproximadamente 0 004 Nótese que el resultado es el mismo que se obtuvo con la prueba Gehan.

A continuación se presenta la prueba Log Rank.

- Prueba Log Rank

La prueba Log Rank está basada en un conjunto de puntuaciones w_i asignados a las observaciones Los puntajes son funciones del logaritmo de la función de sobrevivencia Se estima la función $\log_e$ de supervivencia en el tiempo $t_{(i)}$ utilizando,

$$-e(t_{(i)}) = -\sum_{j \le t_{(i)}} \frac{m_{(j)}}{r_{(j)}} \tag{31}$$

donde $m_{(i)}$ y $r_{(j)}$ fueron definidos en la prueba de Cox-Mantel. Los puntajes sugeridos son $w_{(i)} = 1 - e(t_{(i)})$ para una observación no censurada $t_{(i)}$ y $-e(T)$ para una observación censurada en T. En la práctica, para una observación censurada t_i^+, $w_i = -e(t_{(j)})$, donde $t_{(j)}$ es la observación no censurada más grande tal que $t_{(j)} \le t_i^+$. Entonces, la observación no censurada más grande es la más pequeña en puntaje. Las observaciones censuradas reciben puntajes negativos. La suma de los puntajes w es idénticamente 0 para los dos grupos juntos. La prueba de Log Rank está basada en la suma de S de los w puntajes en uno de los dos grupos La varianza permutacional de S esta dada por,

$$Var(S) = \frac{n_1 n_2 \sum_{i=1}^{n_1+n_2} w_i^2}{(n_1+n_2)(n_1+n_2-1)} \tag{32}$$

que puede ser escrita como,

$$V = \left\{ \sum_{j=1}^{k} \frac{m_{(j)}(r_{(j)} - m_{(j)})}{r_{(j)}} \right\} \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)} \tag{33}$$

La prueba estadística $L = \frac{S}{Var(S)}$ tiene una distribución asintótica normal estándar bajo la hipótesis nula Si S se obtiene del grupo 1, la región crítica es $L < Z_a$ y si S se obtiene del grupo

2, la región crítica se $L > Z_{\alpha}$, donde α es el nivel de significancia para probar $H_o : S_1 = S_2$ contra $H_1 : S_1 > S_2$. El siguiente ejemplo ilustra los cálculos

Ejemplo 5

Considere los datos e hipótesis del ejemplo 3 y 4 utilizados para ilustrar la Prueba Gehan y la Prueba Mantel Cox El estadístico de la prueba Log Rank puede ser calculado por medio de la tabulación $m_{(i)}, r_{(i)}$, $m_{(i)} / r_{(i)}$ y $e(t_{(i)})$ como se ilustra en la siguiente tabla 4. Dado que cada una de las observaciones en las dos muestras, censuradas o no, se les asigna un valor, es conveniente listarlas en la columna 1. De la columna 1 a la 5 solo se encuentran tiempos de fallecimiento; $e(t_{(i)})$ es el valor acumulado de $m_{(i)} / r_{(i)}$. Una estimación (Altshuler, 1970) es el logaritmo natural de la función de sobrevivencia multiplicada por –1.

Tabla 4. Calculo de la Prueba Log Rank

Tiempos de remisión en ambas muestras t_i	$m_{(i)}$	$r_{(i)}$	$m_{(i)} / r_{(i)}$	$e(t_{(i)})$	w_i
15	1	10	.100	.100	0 990[a]
16+	-	-	-	-	- 100
18	1	8	.125	225	775[a]
18+	-	-	-	-	- 225
19	-	6	.333	558	.442[a]
20	2	4	.250	808	192[a]
20+	1	-	-	-	- 808
23	-	2	500	1 308	-1.308
24+	1	-	-		-1.308

a De la muestra 2

Por ejemplo, en $t_{(i)} = 18$, $e(t_{(i)}) = 0.100 + 0.125 = 0.225$; en $t_{(i)} = 19$, $e(t_{(i)}) = 0.225 + 0.333 = 0.558$ 0 La última columna, w_i, proporciona el puntaje para cada observación. Para una observación no censurada $w_i = 1 - e(t_{(i)})$, por ejemplo, en $t_{(i)} = 18$, $w_i = 1 - 0.225 = 0$ 775. Dado que $e(t_{(i)})$ es una estimación de la función de sobrevivencia la cual asume ser constante en dos fallecimientos consecutivos, $e(t_i^+)$ es igual a $e(t_{(j)})$ para $t_{(j)} \leq t_i^+$. Entonces w_i para observaciones censuradas t_i^+ es igual a $-e(t_{(j)})$ donde $t_{(j)} \leq t_i^+$. Por ejemplo, w_i para 16^+ es $-e(15)$, ó –0 100, y que para 18^+ es $-e(18)$, ó –0.225. Observaciones empatadas como

los dos 19's reciben el mismo puntaje:0.442. La suma de las w_i es igualmente cero, lo que puede ser utilizado para verificar los cálculos

El estadístico S es S= 0.900 + 0 775 + 0 442 + 0.442 + 0.192 = 2.751. La varianza de S se calcula con (32) La prueba estadística $L = \frac{2.751}{1.210} = 2.5$, y el valor de p es aproximadamente de 0.0064 Entonces, la hipótesis nula es rechazada al nivel de 0.0064. Los datos mostrados por el tratamiento CMF son superiores

El estadístico Log Rank S puede ser expresado como la suma de los fallecimientos observados menos los fallecimientos condicionales esperados calculados en cada uno de los tiempos de fallecimiento, o como la diferencia entre los fallecimientos observados y esperados en uno de los grupos. Una versión similar de la prueba Log Rank es la prueba Ji-Cuadrada que compara en la hipótesis el número observado de fallecimientos con el número esperado de fallecimientos. Sea O_1 y O_2 los números observados y E_1 y E_2 los números esperados de muertes en los dos grupos de tratamiento. La prueba estadística que a continuación se presenta se distribuye aproximadamente como una Ji-Cuadrada con un grado de libertad.

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \qquad (34)$$

Un valor alto Ji-cuadrado (por ejemplo $\geq \chi^2_{1,.05}$) permitiría rechazar la hipótesis nula a favor de la alternativa de que dos tratamientos no son igualmente efectivos ($\alpha = 0.05$)

Para calcular E_1 y E_2, se arreglan todas las observaciones no censuradas en orden ascendente y se calcula el número esperado de muertes en cada tiempo no censurado y se suman. El número esperado de muertes en un tiempo no censurado es obtenido multiplicando las muertes observadas en ese tiempo por la proporción de pacientes expuestos al riesgo en el grupo de cada tratamiento Sea d_i el número de muertes al tiempo t y n_{1t} y n_{2t} los números de pacientes aún expuesto al riesgo de morir después del tiempo t en los dos grupos de tratamiento Las muertes esperadas para los grupo 1 y 2 al tiempo t son

$$e_{1t} = \frac{n_{1t}}{n_{1t} + n_{2t}} \times d_t \qquad e_{2t} = \frac{n_{2t}}{n_{1t} + n_{2t}} \times d_t$$

Entonces el número total de muertes esperadas en los dos grupos son,

$$E_1 = \sum_{\text{todas las t}} e_{1t} \qquad E_2 = \sum_{\text{todas las t}} e_{2t}$$

En la práctica, solo se necesita calcular el número total de muertes esperadas en uno de los dos grupos, por ejemplo, E_1, dado que E_2 es el número total de muertes menos E_1

En el siguiente ejemplo 6 se ilustran dichos cálculos.

Ejemplo 6

Consideremos los mismo datos hipotéticos utilizados para la prueba Gehan y Mantel Cox.

CMF (grupo 1) 23,16+,18+,20+,24+

Control (grupo 2) 15,18,19,19,20

La hipótesis nula y la alternativa es

$H_0 : S_1 = S_2$ (dos tratamientos igualmente efectivos)

$H_1 : S_1 \neq S_2$ (dos tratamientos no son igualmente efectivos)

La siguiente tabla 5 muestra el calculo de E_1. Por ejemplo, en t=18, cuatro pacientes en el grupo 1 y cuatro en el grupo 2 están aún expuestos al riesgo de recaída , y solo hay una recaída. Entonces, $d_t = 1$, $n_{1t} = n_{2t} = 4$ y $e_{1t} = 0.5$

Tabla 5. Calculo de E_1 para la prueba Log Rank

Tiempo de recaida t	d_t	n_{1t}	n_{2t}	e_{1t}	e_{2t}
15	1	5	5	0.5	0.5
18	1	4	4	0 5	0 5
19	2	3	3	1.0	1 0
20	1	3	1	0.75	0.25
23	1	2	0	1 0	0
Total				3 75	2.25

El número total de recaídas esperadas es $E_1 = 3.75$. Dado que el hay una total de 6 recaídas $(O_1 = 1, O_2 = 5)$ en los dos grupos, $E_2 = 6 - 3.75 = 2.25$. Usando (34), tenemos

$$X^2 = \frac{(1-3.75)^2}{3.75} + \frac{(5-2.25)^2}{2.25} = 5\,378$$

El p valor correspondiente a este valor Ji-cuadrado es menor que 0.05 ($p \cong 0.02$) Entonces, se obtiene la misma conclusión, que en la prueba Gehan y Mantel-Cox, de que hay diferencias significativas entre la efectividad de los tratamientos.

3. Prueba Mantel y Haenszel.

Esta prueba es particularmente utilizada en la comparación de la experiencia de sobrevivencia entre dos grupos cuando son necesarios ajustes con otros factores La prueba ha sido utilizada en varias pruebas clínicas y en estudios epidemiológicos como un método de control de efectos de variables "confusoras". Por ejemplo, comparando dos tratamientos de melanonas malignos, debe ser importante ajustar la comparación por una posible variable "confusora" en alguna etapa de la enfermedad Estudiando la asociación del cigarro en las enfermedades de corazón, resultaría interesante controlar los efectos de la edad Para usar la prueba de Mantel-Haenszel, los datos son estratificados por la variable "confusa" y puestas en una sucesión de tablas de 2X2, una para cada estrato.

Sea s el número de estratos, n_{ji} el número de individuos en el grupo j, $j = 1,2$, y estratos i, $i = 1,...,s$, y d_{ji} el número de muertes (o fallecimientos) en el grupo j y el estrato i Para cada estrato s, la información puede ser representado por una tabla de contingencia 2X2,

	Número de muertes	Número de sobrevivientes	Total
Grupo 1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
Grupo 2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	D_i	S_i	T_i

La hipótesis nula puede ser escrita como,

$$H_0 : p_{11} = p_{12}$$

$$p_{21} = p_{22}$$

$$\vdots$$

$$p_{s1} = p_{s2}$$

donde $p_{ij} = P$ (muerte | grupo j, estrato i). Entonces, la prueba permite simultáneamente la comparación en todas las s tablas de contingencia de las diferencias en sobrevivencia o probabilidades de muerte para los dos grupos

La prueba Ji-Cuadrada sin corrección de continuidad está dada por,

$$X^2 = \frac{\left(\sum_{i=1}^{3} d_{1i} - \sum_{i=1}^{3} E(d_{1i})\right)^2}{\sum_{i=1}^{3} Var(d_{1i})} \qquad (35)$$

donde

$$E(d_{1i}) = \frac{n_{1i} D_i}{T_i} \qquad (36)$$

$$Var(d_{1i}) = \frac{n_{1i} n_{2i} D_i S_i}{T_i^2 (T_i - 1)} \qquad (37)$$

son la media y la varianza, respectivamente, del número de muertes en el grupo i calculadas condicionalmente en los totales marginales de la tabla de contingencia Este estadístico se distribuye aproximadamente Ji-Cuadrada con un grado de libertad. Entonces, un valor calculado Ji-Cuadrada mayor al valor de tablas Ji-Cuadrada para el nivel de significancia establecido indica diferencias significativas en la sobrevivencia entre los dos grupos

La Generalización de la prueba Mantel -Haenszel es la prueba Tarone-Ware

El siguiente ejemplo 7 ilustra la prueba Mantel-Haenszel

Ejemplo 7.

595 personas participan en un estudio control de la asociación del colesterol en enfermedades del corazón (CHD). Entre ellas, 300 personas tienen CHD y 295 no padecen CHD Para investigar si un elevado colesterol está significativamente asociado con CHD, el investigador decide controlar los efectos del cigarro Los sujetos de estudio son divididos en dos estratos fumadores y no fumadores.

Las siguientes tabla 6 y 7 presentan los datos

Tabla 6. Fumadores

Colesterol elevado	Con CHD	Sin CHD	Total
Si	120	20	140
No	800	60	140
Total	200	80	280

Tabla 7. No fumadores

Colesterol elevado	Con CHD	Sin CHD	Total
Si	30	60	90
No	70	155	255
Total	100	215	315

Usando (36) y (37), se obtiene

$$E(d_{11}) = \frac{140 * 200}{280} = 100 \qquad E(d_{12}) = \frac{90 * 100}{315} = 28\,571$$

$$Var(d_{11}) = \frac{140 * 140 * 200 * 80}{(280)^2 (280-1)} = 14.337$$

$$Var(d_{12}) = \frac{90 * 225 * 100 * 215}{(315)^2 (315-1)} = 13.974$$

Usando (35) y d_{11}=120, d_{12}=30, se tiene

$$\chi^2 = \frac{(150-128.57)^2}{14.337+13.974} = 16.220$$

El valor χ^2 cual es significativo al nivel 0.001 Entonces, el elevado colesterol es asociado significativamente con CHD después de ajustar los efectos del cigarro

La generalización de la prueba Mantel y Haenszel se conoce como la prueba Tarone-Ware La hipótesis nula de la prueba Tarone Ware es,

$$H_0 : S_1(t) = S_2(t) = ... = S_r(t)$$

2.4.1.6. Modelo de Cox para datos de sobrevivencia.

En el apartado 2 4.1.4 se describió el método no paramétrico producto - límite de Kaplan y Meier con el cual se estima la función de sobrevivencia de determinado grupo. También se mencionó que es posible obtener la curva de sobrevivencia para diferentes grupos y después establecer si hay diferencias significativas entre ellas. Una de las limitaciones del método no paramétrico producto - límite de Kaplan y Meier es que no hay indicador que proporcione la importancia de las variables ni el efecto simultáneo de las mismas.

En el ejemplo mencionado en el apartado 2.4.1.1 de los pacientes con úlcera péptica que han seguido un tratamiento y en el que algunos de ellos, antes de comenzar el tratamiento, decidieron abandonar el hábito de fumar. Por un lado, se sospecha que , para un mismo tiempo transcurrido desde la respuesta del tratamiento, la probabilidad de que la sintomatología ulcerosa reaparezca depende del efecto del abandono del tabaco. Por otro lado, se sospecha que dicha probabilidad también depende de cuál haya sido el tiempo de respuesta al tratamiento, así como de otros aspectos relacionados con los hábitos del individuo tales como el consumo de alcohol, café o antiácidos Entonces, suponga que se desea estimar la probabilidad de reaparición de los síntomas en función del tiempo transcurrido desde la respuesta al tratamiento, conocidos el tiempo de respuesta al tratamiento y los distintos hábitos del paciente. Con el método no paramétrico producto - límite de Kaplan y Meier sería posible estimar la función de sobrevivencia para los diferentes grupos de las variables y después compararlas entre sí pudiendo establecer si hay o no diferencias significativas, es decir, se podría comparar la función de sobrevivencia de los fumadores y no fumadores o la función de sobrevivencia entre los que consume mucho, poco o medianamente café. Sin embargo, no es posible hasta ahora determinar cuales son las variables

que más determinan la reaparición de la sintomatología ulcerosa ni los efectos simultáneos del tiempo de respuesta al tratamiento, el abandono o no del hábito del cigarro y el consumo nulo, moderado o en demasía de café o antiácido. Por tanto, es necesario el uso de métodos multivariados tal como el *modelo de funciones de riesgo proporcionales de Cox* que se describe en este apartado 2.4.1.5.

El método de regresión múltiple es una técnica convencional para investigar la relación entre el tiempo de supervivencia y determinadas variables de pronóstico. Sea $x_1, x_2, ..., x_p$ las p posibles variables de pronóstico (covariables o variables explicativas). Para el i-ésimo paciente, los valores observados de las p variables son $x_{1i}, x_{2i}, ..., x_{pi}$. En regresión múltiple, el tiempo de sobrevivencia del i-ésimo paciente, t_i , es la variable independiente El interés radica en la identificación de la relación de t_i o una función de t_i, digamos $w(t_i)$ y $x_{1i}, x_{2i}, ..., x_{pi}$ y que pueda ser expresada como una función de regresión.

$$t_i = f_1(x_{1i}, x_{2i}, ..., x_{pi}) = f_1(x_i)$$

o

$$w(t_i) = f_2(x_{1i}, x_{2i}, ..., x_{pi}) = f_2(x_i)$$

Dos ejemplos de f_1 o f_2 son,

$$f_1(x_i) = B_1 x_{1i} + B_2 x_{2i} + ... + B_p x_{pi}$$

$$f_2(x_i) = \exp(B_1 x_{1i} + B_2 x_{2i} + ... + B_p x_{pi})$$

Los modelos propuestos de riesgo para las distribuciones de sobrevivencia generalmente hacen el supuesto de funciones de riesgo proporcionales. Un modelo de funciones de riesgo proporcionales tiene la propiedad de que individuos diferentes tienen funciones de riesgo proporcionales, es decir, $\frac{h(t \mid x_1)}{h(t \mid x_2)}$; el cociente de las funciones de riesgo para dos individuos con covariables $x_1 = (x_{11}, x_{21}, ..., x_{p1})$ y $x_2 = (x_{12}, x_{22}, ..., x_{p2})$ no varían en el tiempo t Esto implica que dado un conjunto de covariables $x_1, x_2, ..., x_p$ la función de riesgo puede ser escrita como,

$$h(t \mid x) = h_0(t) g(x)$$

donde g(x) es una función de x y $h_0(t)$ puede ser considerado como la base de la función de riesgo de un individuo para el que g(x)=1 Este modelo introducido por Cox (1972) es un modelo general no paramétrico apropiado para el análisis de datos de sobrevivencia con o sin datos censurados El modelo utiliza la función de riesgo como la variable dependiente

Cuando los tiempo de sobrevivencia tienen distribución continua y la posibilidad de empates es ignorada, la función de riesgo es,

$$h(t \mid x) = h_0(t)\exp(B_1 x_1 + B_2 x_2 + ... + B_p x_p)$$

$$h(t \mid x) = h_0(t)\exp(\sum_{j=1}^{p} B_j x_j) \qquad (38)$$

donde $h_0(t)$ es la función de riesgo base de la distribución de sobrevivencia (arbitraria) cuando toda las x variables son ignoradas, es decir, cuando todas las variables x son iguales a cero y las $\beta's$ son coeficientes de regresión. De hecho, $\exp(\sum_{j=1}^{p} B_j x_j)$ puede ser remplazado por cualquier función conocida de las $x_j's$ y $\beta's$ Es claro que el modelo de Cox supone que la función de riesgo del grupo de estudio es proporcional a la función base de la distribución de sobrevivencia. Se puede mostrar que (38) es equivalente a,

$$S(t) = [S_o(t)]^{\exp(\sum_{j=1}^{p} \beta_j x_j)} \qquad (39)$$

(Ver anexo 2)

El uso de (38) puede ser ejemplificado como sigue,

1. Caso de dos muestras

Suponga $p = 1$, es decir, solo hay una variable x y sea x_1 una variable indicadora tal que,

$$x_{1i} = \begin{cases} 0 & \text{si el } i\text{-esimo individuo es de la muestra 0} \\ 1 & \text{si el } i\text{-esimo individuo es de la muestra 1} \end{cases}$$

Entonces, de acuerdo a (38), la función de riesgo de las muestras 0 y 1 son respectivamente $h_0(t)$ y $h_1(t) = h_0(t)\exp(\beta_1)$. La función de riesgo de la muestra 1 es igual a la función de riesgo de la muestra 0 multiplicada por una constante $\exp(\beta_1)$, o las dos funciones de riesgo son proporcionales En términos de la función de sobrevivencia,

$$S_1(t) = [S_0(t)]^c$$

donde la constate c= $\exp(\beta_1)$

2. Caso de dos muestras con covariables o variables explicativas

Las x variables en (38) pueden ser variables indicadoras, tal como x_1 en el caso de dos muestras ilustrado en el número anterior, o variables concomitantes. Teniendo uno o mas variables concomitantes x en (38) es permisible examinar la relación entre dos muestras ajustando la presencia de dichas variables

3. Caso de dos muestras con covariables que dependen del tiempo.

En (38) una o más variables x pueden ser funciones del tiempo. Por ejemplo, suponga que se introduce una variable x_1 que depende del tiempo tal que $x_2 = tx_1$ y que además es una variable concomitante Según (38) la función de riesgo en la muestra 1 es,

$$h_1(t) = h_0(t)\exp(\beta_1 + \beta_2 t)$$
$$h_1(t) = ch_0\exp(\beta_2 t)$$

lo que en la muestra 0 es $h_0(t)$.

4. Caso de regresión.

Dividiendo ambos lados de la expresión (38) entre $h_0(t)$ y calculando después el logaritmo natural se obtiene,

$$\frac{h(t \mid x)}{h_0(t)} = \exp(B_1 x_1 + B_2 x_2 + ... + B_p x_p)$$

$$\ln\left[\frac{h(t \mid x)}{h_0(t)}\right] = B_1 x_1 + B_2 x_2 + ... + B_p x_p \qquad (40)$$

El lado izquierdo de la igualdad (40) es una función de riesgo (o de riesgo relativo) del i-ésimo paciente, y el lado derecho es la combinación lineal de las variable concomitantes $x_{1i}, x_{2i}, ..., x_{pi}$ con coeficientes $\beta_1, ..., \beta_p$, respectivamente. Las $x's$ pueden ser variables indicadoras, covariables y/o covariables que dependen del tiempo Si $y_i = \ln\left[\frac{h_i(t)}{h_0(t)}\right]$ entonces y_i es simplemente una ecuación estándar de la ecuación de regresión múltiple con variables concomitantes como variables independientes y una función de riesgo como variable dependiente

El principal interés radica en identificar factores importantes de pronóstico. En otras palabras, se desea identificar de las p variables independientes, un subconjunto de variables relacionadas significativamente con la variable dependiente, y consecuentemente, con el tiempo de sobrevivencia de cada uno de los pacientes. Por tanto, el problema radica en los coeficientes de regresión Si B_i es cero, entonces las variables independientes correspondientes no están relacionadas con la sobrevivencia ajustadas por otras variables independientes Es importante resaltar que en un problema de regresión múltiple asignar rangos a las variables con el propósito de considerar importancia relativa de estas puede ser alcanzado usando un método de regresión paso a paso (stepwise). Por medio de la asignación de rangos y la prueba de significancia para cada variable, se puede seleccionar las variables más significativas en relación con la variable dependiente. Dada la analogía de la regresión múltiple, la regresión paso a paso (stepwise) puede también ser aplicada.

Además, para identificar los factores de pronóstico el modelo de regresión de Cox puede también definir un índice de pronóstico conocido como $\ln\left[\frac{h_i(t)}{h_0(t)}\right]$ para cada uno de los pacientes. Este índice puede ser utilizado para comparar dos grupos de tratamiento así como también el

pronóstico del curso de la enfermedad entre pacientes. Como ya se menciono, $h_0(t)$ es la función de riesgo cuando todas las variables independientes son supuestas igual a cero Si las variables independientes son estandarizadas por la media, el modelo usado es,

$$\ln\left[\frac{h_i(t)}{h_0(t)}\right] = B_1(x_1 - x_1) + B_2(x_2 - x_2) + \ldots + B_p(x_p - \bar{x}_p) \qquad (41)$$

donde x_j es el promedio de la j-ésima variable independiente para todos los pacientes Entonces $h_0(t)$ es la función de riesgo cuando todas las variables toman su valor promedio El índice de riesgo es la razón del riesgo a muerte para un paciente con determinados valores en el conjunto $x_{1i}, x_{2i}, \ldots, x_{pi}$ y un paciente promedio que tiene valores promedio para cada una de las variables. El modelo expresado en (41) es más real y más fácil de interpretar que el modelo expresado en (40) Este índice o razón puede ser utilizado para comparar el riego relativo para pacientes con diferentes valores en las variables independientes

Para estimar los coeficientes $\beta_1, \ldots, \beta_p$ Cox sugiere un procedimiento de máxima verosimilitud donde la función de verosimilitud esta basada en la probabilidad condicional de fallecer. Suponga que $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ son los k tiempos exactos de fallecimiento Sea $R(t_{(i)})$ el conjunto de riesgos al tiempo $t_{(i)}$. $R(t_{(i)})$ consiste en todos los individuos para los cuales los tiempos de sobrevivencia son al menos $t_{(i)}$. Para el tiempo particular de fallecimiento en el tiempo $t_{(i)}$, condicionados en el conjunto de riesgos $R(t_{(i)})$, la probabilidad de fallecer de un individuo se observa como,

$$\frac{\exp(\sum_{j=1}^{p} \beta_j x_{ji})}{\sum_{l \in R(t_{(i)})} \exp(\sum_{j=1}^{p} \beta_j x_{jl})} \qquad (42)$$

Cada fallecimiento contribuye en un factor y el logaritmo natural de la función condicional de máxima verosimilitud es,

$$LL(\beta) = \sum_{i=1}^{k}\sum_{j=1}^{p}\beta_j x_{ji} - \sum_{i=1}^{k}\ln\left[\exp\left(\sum_{j=1}^{p}\beta_j x_{ji}\right)\right] \qquad (43)$$

El estimador de máxima verosimilitud de las $\beta's$ se obtienen resolviendo simultáneamente la p ecuaciones resultantes de derivar $LL(\beta)$ con respecto a $\beta_1,...,\beta_p$ e igualándolas a cero Las p ecuaciones son,

$$U(\beta_1,...,\beta_p) = \sum_{i=1}^{k}\left[x_{ui} - A_{ui}(\beta_1,...,\beta_p)\right] = 0 \qquad u = 1,...,p \qquad (44)$$

donde,

$$A_{ui}(\beta_1,...,\beta_p) = \frac{\sum_{l\in R(t_{(i)})} x_{ul}\exp\left(\sum_{j=1}^{p}\beta_j x_{jl}\right)}{\sum_{l\in R(t_{(i)})}\exp\left(\sum_{j=1}^{p}\beta_j x_{jl}\right)} \qquad (45)$$

Las p ecuaciones de (44) pueden ser resueltas numéricamente por el método de Iteración de Newton-Raphson en el cual los estimadores de $\beta_1,...,\beta_p$ son obtenidos usando iterativamente (44) y la segunda derivada de (43)

$$I_{uv}(\beta_1,..,\beta_p) = -\sum_{i=1}^{k} C_{(uvi)}(\beta_1,..,\beta_p) \qquad (46)$$

donde,

$$C_{(uvi)}(\beta_1,...,\beta_p) = \frac{\sum_{l\in R(t_{(i)})} x_{ul}x_{vl}\exp\left(\sum_{j=1}^{p}\beta_j x_{jl}\right)}{\sum_{l\in R(t_{(i)})}\exp\left(\sum_{j=1}^{p}\beta_j x_{jl}\right)} - A_{ul}(\beta_1,..,\beta_p)A_{vl}(\beta_1,..,\beta_p) \qquad (47)$$

Sin embargo, (38) supone continuidad en los tiempos de sobrevivencia, lo cual no es práctico ya que en la práctica en los tiempos de sobrevivencia frecuentemente hay empates Para cubrir esta posibilidad Cox generaliza (38) y discretiza el tiempo de sobrevivencia por medio de una transformación logística,

$$\frac{h(t)dt}{1-h(t)dt} = \frac{h_0(t)dt}{1-h(t)dt}\exp\left(\sum_{j=1}^{p}\beta_j x_{ji}\right)$$

Suponga que entre los tiempos de sobrevivencia $t_1,...,t_n$, hay k distintos tiempos Sea $t_{(1)},...,t_{(k)}$ los k distintos tiempos de fallecimientos (observaciones no censuradas) Sea $m_{(i)}$ la multiplicidad de $t_{(i)}$, $m_{(i)}$>1, si hay más que una observación con valor $t_{(i)}$, $m_{(i)}$=1 si solo hay una observación con valor en $t_{(i)}$. Sea $R(t_{(i)})$ el conjunto de individuos en riesgo al tiempo $t_{(i)}$ $R(t_{(i)})$ consiste en aquellos individuos para los cuales los tiempos de sobrevivencia son al menos $t_{(i)}$. Sea $r_{(i)}$ el número de tales individuos. Al tiempo $t_{(i)}$, la probabilidad de que el individuo observado fallezca condicionalmente en el conjunto de riesgo $R(t_{(i)})$, de (38)

$$\frac{\exp(\beta_1 z_{1i} + \beta_2 z_{2i} + ... + \beta_p z_{pi})}{\sum_{l\in R(t_{(i)})}\exp(\beta_1 z_{1l} + \beta_2 z_{2l} + ... + \beta_p z_{pl})}$$

donde z_{1i} es la suma de las x_{1i}'s a través de $m_{(i)}$ individuos falleciendo en $t_{(i)}$, z_{2i} es la suma de las x_{2i}'s a través de $m_{(i)}$ individuos falleciendo en $t_{(i)}$, y así sucesivamente El logaritmo natural de la función condicional de máxima verosimilitud es entonces,

$$LL(\beta) = \sum_{i=1}^{k}(\beta_1 z_{1i} + ... + \beta_p z_{pi}) - \sum_{i=1}^{k}\log\left(\sum_{l\in R(t_{(i)})}\exp(\beta_1 z_{1l} + \beta_2 z_{2l} + ... + \beta_p z_{pl})\right) \tag{48}$$

Las funciones U y I en (44) y (46) se convierten en,

$$U(\beta_1,...,\beta_p) = \sum\left[z_{ui} - m_{(i)}A_{ui}\right] = 0 \qquad u = 1,...,p \tag{49}$$

y

$$I_{uv}(\beta_1,...,\beta_p) = \frac{m_{(i)}\left[r_{(i)} - m_{(i)}\right]}{r_{(i)} - 1}C_{(cvi)}(\beta_1,...,\beta_p) \tag{50}$$

Los errores estándar de los estimadores de las β_i's pueden ser estimadas de (46) y (50)

El intervalo de confianza para β_i al 100(1-α)% es,

$\hat{\beta}_i \pm Z_{a_2}$ (Estimador del error estándar de $\hat{\beta}_i$)

Para una variable dicotómica, el modelo de funciones de riesgo proporcionales de Cox puede ser usado para estimar el riesgo ajustado por las otras variables en el modelo Por ejemplo, si x_1 representa la hipertensión y es definida como,

$$x_1 = \begin{cases} 1 & \text{si el paciente es hipertenso} \\ 0 & \text{en otro caso} \end{cases}$$

entonces la tasa de riesgo de los pacientes hipertensos es $\exp(\hat{B}_1)$ veces más alta que los pacientes con presión normal. Esto es, el riego relativo asociado con la hipertensión es $\exp(\hat{B}_1)$ El intervalo de confianza al 100(1-α)% para el riesgo relativo puede ser obtenido usando el intervalo de confianza para β. Sea (β_{1L}, β_{1U}) el intervalo de confianza al 100(1-α)% para β_i; el intervalo de confianza al 100 (1-α)% para el riesgo relativo es $(\exp(\beta_{1L}), \exp(\beta_{1U}))$. Esta aplicación del modelo de funciones de riesgo proporcionales ha sido muy utilizado particularmente por epidemiólogos.

En la estimación de $\beta_1, ..., \beta_p$ debe ser usado un procedimiento paso a paso (stepwise) para asignar rangos a la variable. En el procedimiento hacia delante paso a paso (forward-stepwise), las variables independientes entran en la ecuación de regresión una a una, hasta que la regresión es satisfactoria, es decir, hasta que la última variable ingresada ya no aporte información significativa a las ya seleccionadas. El orden de la inserción de variables es determinada en el uso, por ejemplo, el valor máximo del logaritmo natural de la función de verosimilitud, $LL(\beta)$ [1], es una medida de importancia de las variables que aún no están en la ecuación de regresión. Usando

[1] $LL(\hat{\beta}_i)$ por sus siglas en inglés LogLikelihood

el valor máximo de $LL(\hat{\beta}_i)$ como medida, ella selecciona, como la primera variable para la ecuación de regresión, digamos $x_{(1)}$, aquélla cuyo logaritmo natural de máxima verosimilitud es el valor más alto Sea $LL(\hat{\beta}_i)$, i= 1, . , p, el máximo valor del logaritmo natural de la función de verosimilitud obtenido de ajustar únicamente la i-ésima variable de pronóstico. Entonces $x_{(1)}$ es la primera variable que ingresa a la regresión si,

$$LL(\hat{\beta}_1) = \max_i \left[LL(\hat{\beta}_i)\right]$$

Entonces ahora hay p-1 variable de pronóstico que aún no están ajustadas El máximo valor del logaritmo de la función de máxima verosimilitud $LL\left(\hat{\beta}_{(1)}, \hat{\beta}_{(i)}\right)$ es calculada para cada p-1 variables independientes y la que proporciona el valor más alto del logaritmo de la función de verosimilitud es la siguiente variable que ingresa a la ecuación de regresión El procedimiento continua para ajustar una variable independiente adicional hasta el tiempo en el que la regresión es satisfactoria En cada etapa la prueba de la razón de verosimilitud se desarrolla para determinar si la última variable ingresada aporta información significativa a las ya seleccionadas

En el primer paso, solo hay una variable en la ecuación de regresión que es,

$$\ln \frac{h(t)}{h_0(t)} = \hat{\beta}_{(1)} x_{(1)}$$

donde $x_{(1)}$, la variable más importante relacionada y que puede ser cualquiera de las $x_1, \ldots, x_p$. Para probar la significancia de $x_{(1)}$ se prueba la hipótesis $H_0 : \beta_{(1)} = 0$ utilizando el estadístico,

$$\chi^2 = \frac{\left[U(\beta_{(1)})\right]^2}{I(\beta_{(1)})} \qquad (51)$$

donde U e I se conocen por (49) y (50) Además (51) se distribuye Ji - Cuadrada con un grado de libertad o se distribuye normalmente con media 0 y varianza 1.

$$Z = \frac{[U(\beta_{(1)})]}{I(\beta_{(1)})} \tag{52}$$

Un valor de Ji-cuadrada (o Z) más alto que el 100α % de la distribución Ji - cuadrada con un grado de libertad (o la distribución normal estándar) indica que $x_{(1)}$ es significativamente relacionada a la sobrevivencia al nivel α

En los pasos subsecuentes, la prueba de verosimilitud se lleva a cabo del siguiente modo Sea $LL(\beta_{(1)},...,\beta_{(k)})$ el valor del logaritmo natural de la función de verosimilitud en (43) en el k-ésimo paso después de haber ajustado k variables. La significancia de la k-ésima variable es probada considerando

$$\chi^2 = -2\left[LL(\beta_{(1)},\cdots,\beta_{(k-1)}) - LL(\beta_{(1)},\cdots,\beta_{(k)})\right] \tag{53}$$

la cual se distribuye Ji – Cuadrado con un grado de libertad Un valor Ji – Cuadrada que excede del punto 100α % de la distribución Ji – Cuadrada con un grado de libertad indica que la k-ésima variable que entra a la regresión es significativa al nivel α.

En este procedimiento la primera variable seleccionada es la variable sola más importante en la predicción del tiempo de sobrevivencia, la segunda variable que entre es la segunda más importante y así sucesivamente. Este proceso proporciona una selección sucesiva y la asignación de rangos a las variables independientes de acuerdo a su importancia relativa

El procedimiento de selección hacia delante (forward) es solo una de las posibles maneras de seleccionar que podría resultar en la asignación de rangos a las variables independientes. Otros métodos son el proceso de eliminación hacia atrás (backward) y el procedimiento paso a paso (stepwise) En el procedimiento hacia atrás, se realiza primero la regresión con las p variables y después las variables van siendo eliminadas de la regresión hasta que la regresión es significativa El procedimiento paso a paso (stepwise) permite a la variable que entra a la regresión ser removida en un paso posterior si se encuentra que ya no es importante.

2.4.2. Análisis de conglomerados

2.4.2.1. ¿Qué es el análisis de conglomerados ?

El término análisis de conglomerados representa de forma genérica a un grupo de técnicas de análisis multivariado cuyo propósito es agrupar objetos en grupos o conglomerados, basándose en su semejanza de una serie de atributos. Es un procedimiento que comienza con un conjunto de datos que contiene información sobre una serie de objetos o unidades de análisis e intenta reorganizar estos objetos en un número reducido de grupos, formados por objetos relativamente homogéneos.

Es deseable que los conglomerados resultantes del proceso de análisis muestren alta homogeneidad interna (intra-conglomerado) y alta heterogeneidad externa (entre conglomerados).

Es importante resaltar que para la selección de las variables utilizadas para caracterizar los objetos deberán tenerse en cuenta consideraciones derivadas de la teoría, ya que los resultados pueden verse afectados por la inclusión de variables inadecuadas e irrelevantes, siendo precisamente la elección de variables uno de los aspectos más críticos de la técnica.

La similaridad entre objetos es el punto de partida de cualquier clasificación en el análisis de conglomerados, ya que es la que proporciona las matrices de similaridades (matrices $n \times n$, cuyas filas y columnas son los objetos) En sentido amplio no es más que una medida de la correspondencia o parecido entre los objetos. En el análisis de conglomerados los objetos son evaluados respecto a una serie de características o dimensiones, y éstas deben ser combinadas en una medida de similaridad calculada para todos los pares de objetos, de la misma que se obtienen las correlaciones entre variables.

Una medida de similaridad que mide la cercanía de dos objetos x_i y x_j , es una función d que mapea $R^P \ X \ R^P \rightarrow R^1$ y satisface los siguientes axiomas:

1. $d(i,j) \geq 0$ para toda $i, j \in R^P$
2. $d(i,i) = 0$
3. $d(i,j) = d(j,i)$ para toda $i, j \in R^P$

Si una medida de similaridad además cumple con las condiciones

4. $d(i,j) \leq d(i,k) + d(k,j)$ para toda $i, j \in R^p$

5 $d(i,j) = 0$ si y solo sí $i = j$

se le llama métrica.

A continuación se presentan algunas de las medidas de similaridad utilizadas con mayor frecuencia en el análisis de conglomerados.

1. Distancia métrica:

Un medida sencilla de similaridad es la distancia euclidiana estándar , también llamada *distancia métrica* , que es la distancia entre dos observaciones. La distancia euclidiana se calcula mediante la siguiente ecuación:

Sean x_r y x_s dos puntos entonces la distancia métrica entre x_r y x_s es,

$$d_{rs} = [(x_r - x_s)'(x_r - x_s)]^{1/2}$$

La distancia euclidiana es sensible a las escalas de medición. Una alternativa es utilizar la distancia métrica estandarizada.

2. Distancia métrica estandarizada:

Otra posibilidad para medir la distancia entre una pareja de puntos es, en primer lugar, estandarizar todas las variables y, enseguida, calcula la distancia euclidiana estándar entre los puntos, usando sus valores Z estandarizados. Rara la mayoría de las situaciones, probablemente ésta sea la mejor elección para medir similaridades. La distancia métrica estandarizada se calcula por medio de la siguiente ecuación

Sean z_r y z_s los valores estandarizados de dos puntos entonces la distancia métrica entre z_r y z_s es,

$$d_{rs} = [(z_r - z_s)'(z_r - z_s)]^{1/2}$$

3. Métrica de Mahalanobis

Esta métrica se utiliza para resolver no solo problemas de escalamiento sino también efectos de correlación entre las variables.

$$d_{ij} = \left\{(x_i - x_j)' S^{-1} (x_i - x_j)\right\}^{1/2}$$

donde $\hat{S} = \frac{\sum_m (x_m - \bar{x})(x_m - \bar{x})'}{n-1}$ puede ser una estimación razonable de S que es la matriz de varianzas y covarianzas.

La métrica de Mahalanobis es invariante bajo transformaciones de la forma $y_m = Ax_m + b$ donde A es una matriz no singular.

4. Métrica de Canberra

Es utilizada para variables positivas y generalmente insensible a sesgos y valores distantes

$$d_{ij} = \frac{1}{p} \sum_{k=1}^{p} \left\{ \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}} \right\}$$

5. Métrica de Gower.

En esta métrica el escalamiento a cada variable se da a través de su rango Considera un promedio de los rangos estandarizados de las variables.

$$d_{ij} = \frac{1}{p} \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{R_k}$$

donde R_j es el rango de la variable j.

Una vez calculadas las matrices de similaridad comienza el proceso de particion o formación de grupos. En la práctica existen dos métodos de agrupación

1. Métodos jerárquicos
2. Métodos no jerárquicos o iterativos.

2.4.2.2. Métodos jerárquicos

Los métodos jerárquicos suponen la construcción de una jerarquía de estímulos en forma de árbol Una característica importante de los procedimientos jerárquicos es que los resultados de un temprano estadío están siempre anidados dentro de otro posterior, lo que configura su estructura arborescente, denominada *dendrograma*

Los métodos jerárquicos se subdividen en *aglomerativos y disociativos*. Cada una de estas categorías presenta una gran variedad de métodos.

Los métodos aglomerativos, también conocidas como ascendentes, empiezan el análisis con tantos grupos como individuos haya. A partir de estas unidades iniciales, se van formando grupos de forma ascendente, agrupando cada vez más individuos en los sucesivos grupos que se van formando. Al final del proceso todos los caso están englobados en un mismo conglomerado

Los métodos disociativos, también denominado descendentes o divisivos, constituyen el proceso inverso al anterior. Empiezan con un conglomerado que engloba todos los individuos A partir de este gran grupo inicial, de forma descendente a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso se tienen tantos grupos como individuos.

Existen diversos métodos jerárquicos aglomerativos y disociativos, sin embargo, ninguno de ellos proporciona una solución óptima a todos los problemas. Esto se debe a que es posible llegar distintos resultados según el método elegido. El buen criterio del investigador y el conocimiento del problema sugieren el método mas adecuado y la solución más correcta

2.4.2.3. Métodos no jerárquicos o iterativos.

Los métodos iterativos o no jerárquicos no construyen estructuras de árbol, sino que asignan los objetos una vez que se determina el número de grupos. La mayor parte de los métodos trabajan con un esquema como el siguiente:

1. Se hace una partición de los datos con un numero especifico de conglomerados y se calculan sus *centroides (centro de los conglomerados)*.
2. Se asigna cada objeto al conglomerado a cuyo centroide se parece más.
3. Se calculan nuevos centroides revisando los casos.
4. Se repiten los pasos 2 y 3 hasta que los objetos no cambien de conglomerado.

Los métodos no jerárquicos tienen como objeto realizar una sola partición de los individuos en k grupos. Esto implica que el investigador debe especificar "a priori" los grupos que deben de ser formados. Esta es, posiblemente, la principal diferencia respecto a los métodos jerárquicos. La asignación de individuos a los grupos se hace mediante algún proceso que optimice el proceso de selección. Otra diferencia es que en que estos métodos se trabaja con la matriz de datos original y no requieren su conversión en una matriz de similaridades.

Uno de los algoritmos más conocidos dentro los métodos no jerárquicos o iterativos es el método de k-*Medias* (k-Means). El *método de k-medias* divide un conjunto de individuos en k conglomerados, de tal forma que al final de proceso cada caso pertenece al conglomerado cuyo centro está más cercano a él. La distancia euclidiana es la mediada utilizada para establecer la proximidad entre cada caso y el centro de su respectivo conglomerado El centro del conglomerado es determinado por la media de los individuos que forman cada variable. Cabe resaltar que el investigador establece el número de grupos, es decir, el valor de k.

2.4.2.3.1. Método no jerárquico de k-Medias

Una vez que los k centros de los conglomerados son elegidos, los objetos restantes son asignados al conglomerado cuyo centroide es el más cercano, utilizando alguna medida de proximidad, generalmente la distancia euclidiana cuadrada. El centroide se recalcula cada vez que el conglomerado recibe un nuevo objeto y el proceso se repite hasta que todos los objetos estén clasificados en uno de los k conglomerados.

La distancia entre el i-ésimo individuo y el m-ésimo conglomerado está dado por la siguiente expresión,

$$d_{i,m} = \sum_{j=1}^{p} \left(\left[x_{i,j} - \bar{x}_{m,j} \right]^2 \right)^{1/2} \qquad (1)$$

donde,

$x_{i,j}$ es el valor del i-ésimo individuo sobre la j-ésima variable $i = 1,2,...,n, \qquad j = 1,2,...,p$

$\bar{x}_{m,j}$ es la media del la j-ésima variable en el m-ésimo conglomerado

Sea $P(n,k)$ la partición que resulta de la asignación de cada uno de los n individuos a uno de los conglomerados $1,2,...,k$ y sea n_m el número de individuos que pertenecen al m-ésimo conglomerado.

El error de la partición se define como.

$$E(P_{(n,k)}) = \sum_{i=1}^{n} d^2_{i,m(i)} \qquad (2)$$

donde $m(i)$ es el conglomerado que contiene el i-ésimo individuo, y $d_{i,m(i)}$ es la distancia euclidiana entre el individuo i y la media del conglomerado donde esta contenido el individuo. En el proceso de agrupamiento se busca una partición con un error pequeño, moviendo individuos de un conglomerado a otro hasta que no se transfiera un individuo que resulte en una reducción en el error.

Ejemplo 1:

Consideremos 3 vitaminas contenidas en 6 tipos de alimento de perro.

Tabla 1. Vitaminas contenidas en 6 tipos de alimentos de perro.

Alimento	Vitamina A	Vitamina B	Vitamina C	Suma (i)
Purina	5	9	20	34
Eukanoba	6	11	2	19
Waltham	4	9	20	29
Pedigree	6	9	46	61
Marca Libre	5	7	1	13
Show Show	3	1	12	16

Supongamos que se desea agrupar en tres conglomerados las seis diferentes marcas de alimentos de perro en función de la cantidad de vitamina A, vitamina B y vitamina C que contienen

Los datos se pueden representar por el elemento $a_{i,j}$ que se refiere al i-ésimo individuo con la k-ésima variable donde $1 \leq i \leq 6$ y $1 \leq j \leq 3$

Las siguientes etapas muestran la formación de los conglomerados con el algoritmo no jerárquico de k-medias.

Etapa 1. Como ya se mencionó, el número de conglomerados en el que se desea agrupar estos seis tipos de alimento para perro es tres Es necesario formar conglomerados iniciales Una manera para considerar al individuo i como parte del m-ésimo conglomerado es calculando para cada individuo $k[(sum(i) - \min)]/(\max - \min) + 1$, donde max y min son los valores máximos y mínimos de $sum(i)$. De este modo se forman los tres siguientes conglomerados iniciales,

Tabla 2. Etapa 1: Conglomerados iniciales

Conglomerado	Elementos
1	Eukanoba, Marca Libre, Show Show
2	Purina, Waltham
3	Pedigree

Etapa 2: Se calcula $\bar{x}_{m,j}$,que es la media de la j-ésima variable, de todos los individuos en el m-ésimo conglomerado. Los valores son los siguientes.

Tabla 3. Etapa 2: Media de la j-ésima variable sobre todos los individuos en el m-ésimo conglomerado.

Conglomerado	Vitamina A	Vitamina B	Vitamina C
Eukanoba, Marca Libre, Show Show	14/3	19/3	5
Purina, Waltham	9/2	7	20
Pedigree	6	9	46

Etapa 3: Se calculan las distancias del i-ésimo individuo al j-ésimo conglomerado según la expresión (1),

El error en la partición es,

$$E(P(6,3)) = d^2_{1,2} + d^2_{2,1} + d^2_{3,2} + d^2_{4,3} + d^2_{5,1} + d^2_{6,1}$$

$$E(P(6,3)) = (5-\frac{9}{2})^2 + (9-7)^2 + (20-20)^2 + (6-\frac{14}{3})^2 + (11-\frac{19}{3})^2 + (2-5)^2 + (4-\frac{9}{2})^2 +$$

$$(20-20)^2 + (6-6)^2 + (9-9)^2 + (46-46)^2 + (5-\frac{14}{3})^2 + (7-\frac{19}{3})^2 + (1-5)^2 + (1-\frac{19}{3})^2 + (12-5)^2$$

$$E(P(6,3)) = 137.805$$

Etapa 4. Se verifica si algún movimiento de un individuo de un conglomerado a otro produce una reducción en E (error) calculando la siguiente expresión para cada uno de los individuos,

$$R_{m(i),m} = \frac{n_m d^2_{i,m}}{n_m + 1} - \frac{n_{m(i)} d^2_{i,m(i)}}{n_{m(i)} + 1} \qquad (3)$$

donde,

$n_{(m)}$ = número de individuos en el m-ésimo conglomerado

$m(i)$ = i-ésimo individuo contenido en el conglomerado

Para el primer individuo tenemos,

$$d^2_{1,1} = (5-\frac{14}{3})^2 + (9-\frac{19}{3})^2 + (20-5)^2 = 232.22$$

$$d^2_{1,2} = (5-\frac{9}{2})^2 + (9-7)^2 + (20-20)^2 = 4.25$$

$$d^2_{1,3} = (5-6)^2 + (9-9)^2 + (20-46)^2 = 677$$

$$R_{2(1),1} = \frac{3}{4}(233.22) - 2(4.25) = 166.66 > 0$$

$$R_{2(1),3} = \frac{677}{2} - 2(4.25) = 330.25 > 0$$

Se puede observar que hay un cambio del primer individuo en el segundo conglomerado o al tercero y produce un incremento en E por lo que el primer individuo permanecerá en el segundo conglomerado

Cálculos similares se hacen para los individuos 2,3,4,5,y 6 En el último caso, el individuo 6, los resultados muestran que hay una reducción en el error, si este individuo se mueve del primer conglomerado al segundo. La reducción es de 52.15 Sin embargo ahora se encuentra que,

$$E(P'_{n,k}) = 137.805 - 52.15 = 85.655$$

donde la partición está conformada por: (Eukanoba, Marca Libre), (Purina, Waltham,Show Show) y (Pedigree).

Etapa 5: Se recalculan los valores de $\bar{x}_{m,j}$, cuyos valores se presentan en la siguiente tabla 4.

Tabla 4. Etapa 5: Valores de $\bar{x}_{m,j}$ recalculados.

Conglomerado	Vitamina A	Vitamina B	Vitamina C
Eukanoba, Marca Libre	11/2	9	3/2
Purina, Waltham, Show Show	4	5	52/3
Pedigree	6	9	46

Se regresa a la etapa 4 y se calcula los valores de $R_{m(i),m}$ que son positivos y por tanto si no hay más cambios se producirán errores más pequeños. Por tanto, los conglomerados finales quedan como,

Tabla 5. Conglomerados finales

Conglomerado	Elementos
1	Eukanoba, Marca Libre
2	Purina, Waltham, Show Show
3	Pedigree

CAPÍTULO 3.
ANÁLISIS DE DATOS

3.1. Estrategia de análisis

El objetivo inicial de este trabajo era estudiar la deserción estudiantil sin embargo la falta de información apropiada sugiere abordar este tema con un término alterno *retención estudiantil* Un alumno es retenido[1] en la universidad si habiéndose matriculado en un periodo lectivo, aparece matriculado en el siguiente

Como ya se menciona en el Capítulo 2, la Universidad Autónoma Metropolitana establece como periodo máximo para el término de los estudios de licenciatura diez años. Luego entonces, para poder determinar si un alumno es desertor sería necesario contar con información relativa a diez años de trayectoria escolar; la población objetivo de este trabajo es la cohorte de estudiantes que ingresaron a la división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana en la primavera y otoño de 1995. Naturalmente no es posible tener información de la trayectoria escolar de dichos estudiantes durante diez años

La retención estudiantil guarda una relación inversa con la deserción estudiantil. Es decir, entre mayor sea la deserción estudiantil en una universidad menor será la retención estudiantil y del mismo modo, a mayor retención estudiantil, la deserción es menor.

Por otro lado resulta importante estudiar el desempeño académico de los estudiantes El desempeño académico es el avance en créditos del plan de estudios durante el tiempo que el estudiante permanece en la universidad. Es decir, un estudiante con un desempeño académico del 100% es un estudiante que cubrió los créditos conforme lo señala el plan de estudios Del mismo modo un estudiante con desempeño académico menor del 100% es aquel que en algún momento de su trayectoria escolar se atrasó en relación a los créditos cubiertos y establecidos en el plan de estudios: este tipo de alumno es un alumno rezagado. El desempeño académico está relacionado con la retención estudiantil. Podría suponerse que a un estudiante con alto desempeño académico tiene menor probabilidad de desertar y por tanto mayor probabilidad de seguir inscrito en la universidad. Este concepto de continuar matriculado en la universidad se le conoce como retención, es decir, un alumno que habiéndose matriculado en un trimestre dado, aparece matriculado en el siguiente es retenido en la universidad. Del mismo modo, un estudiante con bajo desempeño académico tiene mayor probabilidad de desertar y por tanto menor probabilidad de ser retenido en la universidad

[1] Glosario de la Educación Superior, *ANUIES-SEP 1984*

ESTA TESIS NO SALE
DE LA BIBLIOTECA

La estrategia para el análisis de datos consiste en 3 etapas:

Etapa 1. Análisis preliminar: desempeño académico

Caracterizar el desempeño académico por carrera y periodo de ingreso bajo el supuesto de que a mayor desempeño académico menor retención. El propósito de esta etapa es obtener una idea general del desempeño académico de los estudiantes.

Etapa 2. Análisis de conglomerados.

Por medio del análisis de conglomerados se agrupa a los estudiantes de acuerdo a sus características, es decir, género, edad al ingreso, escuela de procedencia de educación media superior, promedio en la educación media superior y el porcentaje de aciertos obtenidos en cada una de las áreas del examen de admisión, es decir, en el área de razonamiento matemático, razonamiento verbal y conocimientos generales. El propósito del análisis de conglomerados es formar grupos de estudiantes con características homogéneas y después, bajo la hipótesis de que las características de ingreso de un estudiante determinan su desempeño académico durante su estancia en la universidad, modelar el desempeño académico de cada uno de los grupos. Nótese que modelar el desempeño académico de cada uno de los grupos compuestos por alumnos con características un tanto homogéneas permitirá asociar las características al nuevo ingreso de un determinado estudiante con su desempeño académico.

En esta parte se utilizará el método no jerárquico de k-medias dado que dicho método utiliza el cuadrado de la distancia euclidiana como medida de similaridad, sólo las variables con escala de razón serán integradas en el análisis, es decir, edad al ingreso, promedio en la educación media superior y el porcentaje de aciertos obtenidos en cada una de las áreas del examen de admisión tales como, razonamiento matemático, razonamiento verbal y conocimientos generales. La escuela de procedencia de educación media superior y el género será integrados posteriormente a la formación de los grupos.

Etapa 3: Análisis de sobrevivencia.

La etapa tres se divide en las dos subetapas complementarias:

1. Método no paramétrico producto-límite de Kaplan y Meier.

El método no paramétrico producto-límite de Kaplan y Meier estima la curva de sobrevivencia para un determinado grupo o grupos. La curva de sobrevivencia determina la probabilidad, en función del tiempo, de que los estudiantes de dicho grupo o grupos sean no retenidos en la universidad. Además, pruebas no paramétricas tales como Log Rank, Tarone Ware y Breslow permiten establecer si hay diferencias significativas entre las curvas de sobrevivencia de los grupos.

El método no paramétrico producto límite de Kaplan y Meier permitirá estimar la función de sobrevivencia de cada uno de los conglomerados y después, por medio de las pruebas no paramétricas ya mencionadas, establecer si hay diferencias significativas entre ellas En principio esto permite estimar la probabilidad de los integrantes de cada uno de los cinco conglomerados de sobrevivir al evento crítico no - retención, es decir, la probabilidad en función del tiempo de que sean retenidos en la universidad. Recordando que cada uno de los conglomerados está integrando por estudiantes con características homogéneas, entonces, dadas ciertas características un alumno es asignado en un conglomerado y después tener la estimación de la probabilidad, en función del tiempo, de sobrevivencia.

Además el modelo no paramétrico producto-límite de Kaplan y Meier se utilizará para las variables género, carrera, edad al ingreso, porcentaje de aciertos en las diferentes áreas del examen de admisión, escuela de procedencia y promedio en el nivel educativo anterior. Las variables de razón se categorizarán en este paso para fines del mismo Esto permitirá comparar las funciones de sobrevivencia de dichos grupos Por ejemplo, se comparará la curva de sobrevivencia de hombres y mujeres determinando la existencia o no de diferencias significativas entre ambas curvas de sobrevivencia..

Los resultados del método producto-límite de Kaplan y Meier son limitados y solo permiten sacar conclusiones parciales ya que hasta ahora no se tiene un modelo multivariado en el que en función de las características de ingreso de los estudiantes se les asigne una probabilidad de ocurrencia del evento crítico no retención, es decir, que la probabilidad de sobrevivencia sea una sobrevivencia sea la variable dependiente y las variables independientes el género, la edad al ingreso, la escuela de procedencia de educación media superior, el promedio en la educación media superior y el porcentaje de aciertos obtenidos en cada una de las áreas del examen de admisión, es decir, en el área de razonamiento matemático, razonamiento verbal y conocimientos generales. Hasta ahora, lo más cercano a un modelo multivariado es el análisis de curvas de sobrevivencia por conglomerado pero recordemos que en el análisis de conglomerados no será posible incluir de primera entrada las variables género y escuela de procedencia y por tanto serán caracterizadas de manera muy general, además que primero se clasificó a los estudiantes después se les caracterizó por su desempeño académico y después se le asignó una probabilidad de sobrevivencia al evento crítico no - retención.

2. Modelo de funciones de riesgo proporcionales 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, el modelo de funciones de riesgo proporcionales consiste en obtener una función lineal de las variables independientes que permita estimar, en función del tiempo, la probabilidad de que ocurra dicho suceso

El modelo de funciones de riesgo proporcionales es un modelo de regresión por tanto se utilizará el método paso a paso hacia adelante (forward) para seleccionar las variables que mejor determinan la probabilidad de que el evento crítico no retención ocurra. En esta etapa todas las variables de interés entran al modelo y es posible medir su efecto multivariado.

3.2. Análisis preliminar de los datos.

Recordemos el supuesto que a mejor desempeño académico, el estudiante tiene mayor probabilidad de permanecer en la Institución y por tanto menor probabilidad de abandonarla Como se menciono en el apartado anterior *3.1. Estrategia de análisis*, el propósito de esta etapa es obtener una idea general del desempeño académico de los estudiantes de las nueve diferentes carreras diferenciando los dos diferentes periodos de ingreso: primavera y otoño En el apartado 3.2.1 se define el desempeño académico, en el 3.2 2. se presenta el análisis para una de las carreras Ingeniería Biomédica (el análisis de las ochos restantes licenciaturas se encuentran en el Anexo 3). Finalmente en el apartado 3.2.3 se presentan los resultados generales obtenidos de las nueve licenciaturas.

3.2.1. Definición de desempeño académico

Se define el desempeño académico como el porcentaje de créditos aprobados por unidad de tiempo. La unidad de tiempo es el trimestre lectivo.

Sea d_t el desempeño académico de un estudiante al tiempo t.

$$d_t = \sum_{t=1}^{t} P_{Cat} \quad t=1,..,12$$

donde P_{Cat} es el porcentaje de créditos, en relación con el total, aprobados en el período t.

El valor mínimo del desempeño académico d_t es cero y el máximo es 100.

Además se definen como,

- *Desempeño académico al 100%:* aprobación del porcentaje del número de créditos establecidos por la universidad en el plan de estudios para cada uno de los 12 trimestres.
- *Deficiencia del desempeño académico al 50%:* aprobación de la mitad del porcentaje de créditos establecidos por la universidad en el plan de estudios para cada uno de los 12 trimestres.

La división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana Unidad Iztapalapa imparte las siguientes nueve licenciaturas:

1. Licenciatura en Ingeniería Biomédica
2. Licenciatura en Ingeniería Hidrológica

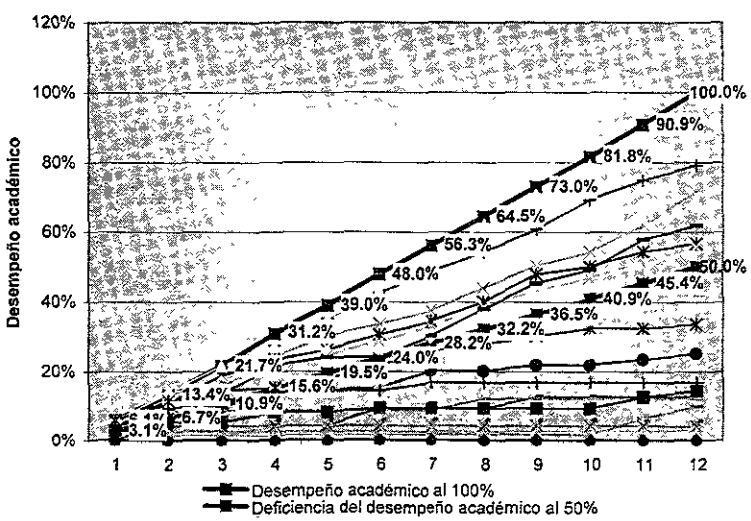
3. Licenciatura en Ingeniería Química
4. Licenciatura en Ingeniería en energía
5. Licenciatura en Física
6. Licenciatura en Ingeniería Electrónica
7. Licenciatura en Matemáticas
8. Licenciatura en Química
9. Licenciatura en Computación

A continuación se analiza el desempeño académico de los estudiantes que ingresaron en el periodo primavera y otoño de 1995 a la carrera de la licenciatura de Ingeniería Biomédica.

3.2.2. Análisis del desempeño académico: Ingeniería Biomédica

En el período primavera 1995 ingresaron 27 estudiantes a la licenciatura de Ingeniería Biomédica. La siguiente gráfica 1 ilustra su trayectoria escolar en función del desempeño académico a lo largo de los 12 trimestres que el plan de estudios establece.

Gráfica 1. Desempeño académico: Licenciatura en Ingeniería Biomédica. Periodo y Año de ingreso: Primavera 1995



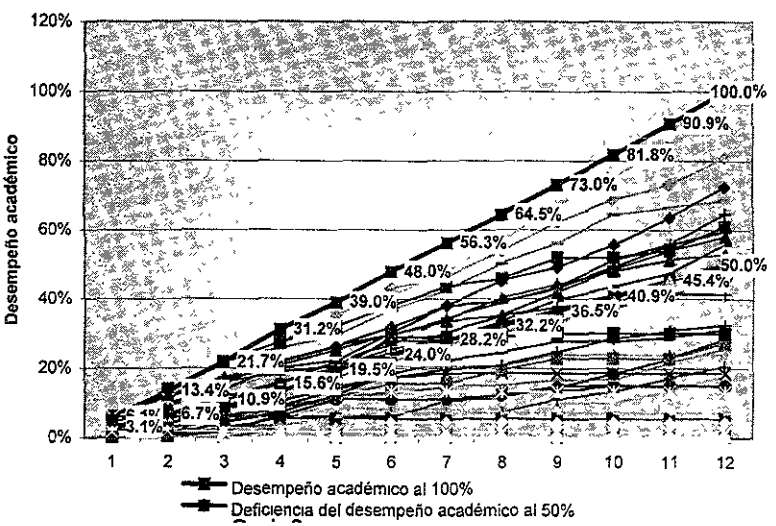
Se observan casos en los que los estudiantes tienen una deficiencia en su desempeño académico menor al 50%; sin embargo, la mayoría de los alumnos presentan una deficiencia

mayor. Además se observan un número significativo de casos en los que la trayectoria académica esta representada por rectas paralelas al *eje x* y que interceptan al *eje y* en el valor correspondiente al desempeño académico en el primer o segundo periodo, esto representa a los alumnos que se estancaron a partir de dichos periodos.

Hay dos casos en los que el desempeño académico a lo largo de los 12 trimestres esta muy cercano al avance establecido por el plan de estudios. La tasa de rezago es del 100% dado que no hay ningún alumno con desempeño académico que cumpla con el avance y tiempo estipulado por la universidad.

Por otro lado, para el caso en que los estudiantes ingresaron en el período Otoño se tienen 56 observaciones cuya trayectoria académica en función de su desempeño académico se ilustra a continuación

Gráfica 2. Desempeño académico: Licenciatura en Ingeniería en Biomédica
Período y Año de ingreso: Otoño 1995



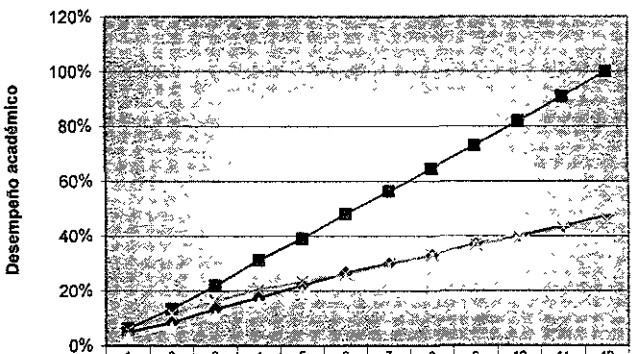
En este caso también se observa que no hay ningún alumno cuyo avance de créditos cumpla con los tiempos establecidos por el plan de estudios.

Se observan algunos casos en los que el desempeño académico está acotado por el desempeño académico al 100% y la deficiencia del desempeño académico al 50%. Sin embargo se observa que la mayoría de estudiantes se encuentran por debajo de la recta que representa la

deficiencia al desempeño académico del 50%. Además no dejan de ser importantes, las líneas paralelas al eje x que representan alumnos que a partir del primer o segundo trimestre, en el mejor de los casos, no contribuyen en su avance de créditos.

La gráfica 3 que a continuación se presenta compara el desempeño académico promedio de los estudiantes que ingresaron en el periodo primavera y los que ingresaron en el periodo otoño. Además está representada la recta de desempeño académico al 100% y la de la deficiencia del desempeño académico al 50%.

Gráfica 3. Desempeño Académico de la Ingeniería en Biomédica
Año y Período de ingreso: Otoño y Primavera de 1995



	1	2	3	4	5	6	7	8	9	10	11	12
Desempeño académico promedio Período Otoño	4.9%	8.4%	13.4%	17.5%	21.9%	26.5%	29.8%	33.1%	36.9%	40.2%	43.5%	47.4%
Desempeño académico al 100%	6.1%	13.4%	21.7%	31.2%	39.0%	48.0%	56.3%	64.5%	73.0%	81.8%	90.9%	100.0%
Deficiencia del desempeño académico al 50%	3.1%	6.7%	10.9%	15.6%	19.5%	24.0%	28.2%	32.2%	36.5%	40.9%	45.4%	50.0%
Desempeño promedio promedio Período primavera	5.6%	11.8%	16.3%	20.5%	23.3%	26.1%	29.7%	33.2%	36.7%	39.4%	42.9%	45.7%

Se observa que la diferencia entre el desempeño académico promedio entre los estudiantes que ingresaron en primavera y los que ingresaron en el periodo otoño es casi nula En los primeros nueve trimestres los alumnos del periodo primavera presentan ligeramente un mejor desempeño, sin embargo al final del doceavo periodo se encuentran por debajo de los que ingresaron en Otoño.

En el Anexo 3 se presenta el análisis preliminar de la variable desempeño académico realizado para las restantes ocho carreras que imparte la división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana Unidad Iztapalapa. A continuación se presentan los resultados generales obtenidos de este análisis.

3.2.3. Resultados generales obtenidos de la variable desempeño académico.

En la siguiente tabla 1 se resume el análisis, por licenciatura y período de ingreso, de la trayectoria académica en función del desempeño académico de los estudiantes durante los 12 trimestres que establece la universidad como plazo regular para el termino de los estudios En ella se presenta el porcentaje de estudiantes y el porcentaje de desempeño académico alcanzado al trimestre 12 En todos los casos se observo una tasa de rezago del 100%

Tabla 1. Desempeño académico por carrera y periodo de ingreso al trimestre 12.

Número	Carrera	Período de ingreso	Total nuevo ingreso	Desempeño académico al trimestre 12 del total de nuevo ingreso			
				Hasta el 25%	Más del 25% hasta el 50%	Más del 50% hasta el 75%	Más del 75%
1	Física	Primavera	19	68.4%	15.8%	15 8%	0.0%
	Física	Otoño	50	90.0%	8 0%	2.0%	0 0%
2	Ingeniería Biomédica	Primavera	27	66.7%	11.1%	14 8%	7 4%
	Ingeniería Biomédica	Otoño	56	58.9%	19 6%	17.9%	3 6%
3	Ingeniería Hidrológica	Primavera	11	81.8%	18.2%	0 0%	0.0%
	Ingeniería Hidrológica	Otoño	15	86.7%	13.3%	0 0%	0.0%
4	Ingeniería Química	Primavera	40	67.5%	27.5%	5.0%	0.0%
	Ingeniería Química	Otoño	47	72.3%	19.1%	6 4%	2.1%
5	Ingeniería en Energía	Primavera	30	83.3%	13 3%	3 3%	0 0%
	Ingeniería en Energía	Otoño	52	73 1%	17 3%	9 6%	0 0%
6	Ingeniería Electrónica	Primavera	58	55.2%	34 5%	3.4%	6 9%
	Ingeniería Electrónica	Otoño	77	39.0%	26.0%	27.3%	7.8%
7	Matemáticas	Primavera	60	78.3%	15 0%	5.0%	1 7%
	Matemáticas	Otoño	75	69.3%	16 0%	14.7%	0 0%
8	Química	Primavera	8	62.5%	25 0%	12.5%	0 0%
	Química	Otoño	35	91 4%	2.9%	5.7%	0 0%
9	Ingeniería en Computación	Primavera	58	51 7%	24.1%	19 0%	5 2%
	Ingeniería en Computación	Otoño	67	56.7%	16 4%	16.4%	10 4%
Total			785	66 4% 521	18 7% 147	11 6% 91	3 3% 26

El 66.4% del total de alumnos que ingresaron en ambos periodos tiene una deficiencia en su desempeño académico del 75%, es decir, han avanzado hasta un total del 25%, al doceavo

trimestre, de lo que establece la universidad en las nueve diferentes carreras. Por otro lado el 11 6% de los estudiantes han alcanzado un desempeño académico entre el 50% y 75% y sólo el 3 3% ha acreditado el 75% de los créditos establecidos en el plan de estudios respectivo

El 100% de los estudiantes son alumnos rezagados son alumnos que por cualquier razón no mantienen el ritmo regular del plan de estudios y su egreso ocurre en una fecha posterior al establecido en dicho plan

La tabla 1 evidencian el bajo desempeño académico de los alumnos de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana. En general, se observa un desempeño académico con mayor deficiencia en los alumnos que ingresaron en el período primavera, esto puede ser explicado por el hecho de que son estudiantes que no ingresaron a la universidad inmediatamente después de finalizar su educación media superior.

Entre las carreras que reflejan una deficiencia mayor en el desempeño académico están Matemáticas, Química, Ingeniería en Energía e Ingeniería Hidrológica y las carreras cuyos estudiantes presentan un mejor desempeño académico son Ingeniería Electrónica e Ingeniería en Computación

Las diferencias observadas en los resultados del análisis del desempeño académico sugieren la necesidad de enfatizar las características que hacen que un estudiante tenga un rendimiento más alto que otro. El análisis de conglomerados, realizado a continuación, permitirá caracterizar grupos de estudiantes de acuerdo a sus características tales como su género, edad al ingreso, escuela de procedencia del nivel educativo anterior, promedio en el nivel educativo anterior y resultados en el examen de admisión diferenciados por habilidad verbal, habilidad matemática y conocimientos generales. Posteriormente se buscará modelar, de manera general, el desempeño académico de cada uno de los grupos.

3.3. Aplicación del análisis de conglomerados: método no jerárquico de K medias.

El análisis de conglomerados busca la segmentación del universo en grupos homogéneos, es una técnica multivariada para detectar agrupación en los datos Los objetos en estos grupos pueden ser casos o variables. El análisis de conglomerados es similar al análisis de discriminante en el sentido de que el investigador busca clasificar un conjunto de objetos en grupos o categorías sin embargo en el análisis de conglomerados se desconoce el número y los miembros de los grupos Es decir, en el análisis de conglomerados , se empieza sin saber la pertenencia a los grupos y regularmente no se sabe cuantos conglomerados hay.

El método no jerárquico de k-medias es utilizado comúnmente en los problemas en los que el tamaño de la población en estudio es grande (200 o más casos). Los métodos jerárquicos calculan una matriz de distancia con entradas para cada par de casos o variables por lo que para tamaño de poblaciones muy grandes se hace difícil su manejo e interpretación.

El método de k-medias empieza usando los valores de los k primeros casos (o variables) como estimación temporal de las medias de los k conglomerados, donde k es el número de conglomerados especificados por el usuario. Los centros iniciales de los conglomerados se forman asignando cada caso (o variable) en el conglomerado con el centro más próximo y después recalculando el centro del conglomerado. Entonces, un proceso iterativo se utiliza para encontrar los centros finales de los conglomerados. En cada paso los casos (o variables) son agrupados en el conglomerado con el centro más próximo y los centros de los conglomerados son recalculados Este proceso continua hasta que no ocurren más cambios en los centros de los conglomerados o hasta que un número determinado de iteraciones es alcanzado.

El método no jerárquico de k medias se utilizará para agrupar a los estudiantes de acuerdo a su edad al ingreso, promedio en el nivel educativo anterior y resultados en el examen de admisión diferenciados por habilidad verbal, habilidad matemática y conocimientos generales Las variables género y escuela de procedencia en el nivel educativo anterior no serán incluidas directamente en el método k medias ya que este método utiliza el cuadrado de la distancia euclidiana como medida de similaridad, lo que exige que las variables sean de escala de razón. Con el objeto de que las variables con valores más altos no contribuyan más en la medida de similaridad que las variables con valores más pequeños, las variables se estandarizan (puntuaciones z) con la siguiente transformación,

$$z_i = \frac{x_i - \bar{x}}{s}$$

donde $\bar{x}$ es la media de la variable x_i y s la desviación estándar de x_i

La justificación de utilizar esta herramienta del análisis de conglomerados radica en las diferencias observadas en el desempeño académico de los estudiantes (apartado 3 2) Resulta necesario establecer las características que determinan que un estudiante tenga mejor desempeño académico que otro bajo el supuesto de que un alumno con buen desempeño académico tiene "mayor potencial de éxito" y por tanto mayor probabilidad de ser retenido en la universidad y menor probabilidad de desertar.

Como ya se mencionó, en el método no jerárquico de k-Medias es necesario especificar el número de grupos o conglomerados en el que se desea agrupar los casos. Se estableció k=5 pensando en que el desempeño académico esta caracterizado por dos grupos en los extremos, es decir, por un lado aquellos estudiantes con mejor desempeño académico y por otro, aquellos estudiantes cuyo desempeño académico durante los 12 trimestres es casi nulo. Además un grupo central y otros dos grupos que se encuentran entre este grupo central y el de mejor desempeño académico y entre este grupo central y el de peor desempeño académico.

La siguiente tabla 1 muestra el tamaño de cada uno de los cinco conglomerados resultantes y en la tabla 2 el desempeño académico de los estudiantes de cada uno de los cinco conglomerados.

(Ver Anexo 4)

Tabla 1. Descripción de los conglomerados.

Número de conglomerado	Tamaño del conglomerado
1	156
2	172
3	210
4	70
5	177
Total	785

Tabla 2. Desempeño académico al trimestre 12 por conglomerado.

		Desempeño académico al doceavo trimestre				Total
		Hasta el 25%	Del 25% hasta el 50%	Del 50% hasta el 75%	Más del 75%	
Conglomerado	1	123 78.8% 23 6%	19 12 2% 13 1%	14 9 0% 15 4%		156 100 0% 19 9%
	2	132 76 7% 25 3%	27 15 7% 18 6%	12 7 0% 13 2%	1 6% 3 6%	172 100 0% 21 9%
	3	130 61 9% 25 0%	45 21 4% 31 0%	28 13.3% 30 8%	7 3.3% 25 0%	210 100.0% 26 8%
	4	48 68 6% 9 2%	17 24.3% 11.7%	3 4.3% 3.3%	2 2 9% 7 1%	70 100 0% 8.9%
	5	88 49.7% 16.9%	37 20 9% 25.5%	34 19.2% 37.4%	18 10 2% 64 3%	177 100 0% 22 5%
Total		521 66 4% 100.0%	145 18 5% 100.0%	91 11 6% 100 0%	28 3 6% 100.0%	785 100 0% 100 0%

En la Tabla 2 se observa que el 64.3% de los estudiantes cuyo desempeño académico al doceavo trimestre es mayor del 75% se encuentran en el conglomerado 5, además en este mismo conglomerado se encuentra el mayor porcentaje (37.4%) de los estudiantes cuyo desempeño académico al doceavo trimestre oscila entre el 50% y 75% . Por otro lado, en el tercer conglomerado, se encuentra el 25% de estudiantes cuyo desempeño académico al doceavo trimestre es del 75% y el 30.8% de aquellos cuyo desempeño académico al doceavo trimestre oscila entre el 50% y 75%. En estos dos conglomerados, tercero y quinto, se concentran el 89 3% de los estudiantes con desempeño académico mayor del 75% y el 68.2% con desempeño académico entre el 50% y 75%.

Los tres restantes conglomerados presentan características muy similares. Sin embargo en el conglomerado cuatro se observa que el porcentaje de estudiantes con desempeño mayor del 75% es mayor que en los conglomerados dos, 2.9% y .6% respectivamente. En el primer conglomerado no hay observaciones con desempeño académico al doceavo trimestre mayor del 75%. Por otro lado, a pesar de que en primer conglomerado, el porcentaje de estudiantes con desempeño académico entre el 50% y 75% (9%) es mayor que en los conglomerados cuarto y

segundo, el primer conglomerado tiene el mayor porcentaje de estudiantes con desempeño menor al 25% que los conglomerados 2 y 4, 78.8%, 76 7% y 68 6% respectivamente.

Es importante señalar que en los 5 conglomerados, el grupo de estudiantes que predomina es el que tiene un desempeño académico menor al 25% Sin embargo, este porcentaje varía para cada uno de los conglomerado siendo el quinto conglomerado donde el grupo es más pequeño (49 7%) y el primer conglomerado quien tiene el mayor porcentaje de estudiantes con esta característica (78 8%).

Entonces, bajo el supuesto de que a mejor desempeño académico mayor probabilidad de retención y por tanto menor probabilidad de deserción, es posible etiquetar a los conglomerados del siguiente modo:

Tabla3. Caracterización de los Conglomerados.

Conglomerado	Renumeración del conglomerado	Tamaño del conglomerado	Etiqueta
5	1	177	Mayor potencial exito
3	2	210	Aceptable potencial de éxito
4	3	70	Moderado potencial de éxito
2	4	172	Bajo potencial de éxito
1	5	156	Casi nulo potencial de éxito
	Total	785	

La tabla 4 presenta los centros finales de los conglomerados (no estandarizados), es decir, el valor promedio de cada una de las variables en cada uno de los cinco conglomerados.

Tabla 4. Centros finales de los conglomerados renumerados no estandarizados

	Conglomerados renumerados				
	1	2	3	4	5
Edad al ingreso	21 56	23.52	32 39	22.58	22 14
Porcentaje de aciertos en razonamiento matemático	73.65	76.78	78 32	54.78	55 62
Porcentaje de aciertos en conocimientos generales	45.52	53.87	60 20	33.16	30 96
Porcentaje de aciertos en razonamiento verbal	66.17	60.89	61 62	38.56	63.20
Promedio en el nivel educativo anterior	8 85	7.50	8.10	7.81	7.68

A continuación se presenta el análisis de varianza de cada una de las variables involucradas en el análisis de conglomerados, es decir, edad al ingreso, promedio en el ciclo anterior y porcentaje de aciertos en las distintas áreas del examen de admisión

Tabla 5. Análisis de varianza (ANOVA) de las variables edad al ingreso, promedio en el ciclo anterior y porcentaje de aciertos en las distintas áreas del examen de admisión.

		Sum of Squares	df	Mean Square	F	Sig.
Edad al ingreso	Between Groups	6639.040	4	1659 760	203.182	000
	Within Groups	6371.679	780	8.169		
	Total	13010.718	784			
Porcentaje de aciertos en razonamiento verbal	Between Groups	83630 909	4	20907 727	173 809	000
	Within Groups	93827 106	780	120 291		
	Total	177458.015	784			
Porcentaje de aciertos en razonamiento matemático	Between Groups	82778 487	4	20694 622	219 034	000
	Within Groups	73695.305	780	94 481		
	Total	156473 792	784			
Porcentaje de aciertos en conocimientos específicos	Between Groups	85742.863	4	21435 716	215.846	000
	Within Groups	77461.928	780	99 310		
	Total	163204.791	784			
Promedio en el nivel educativo anterior	Between Groups	202.306	4	50.576	190 986	000
	Within Groups	206.557	780	265		
	Total	408 863	784			

Como era de esperarse, dado que el análisis de conglomerados busca heterogeneidad entre los grupos , el análisis de varianza señala en todos lo casos diferencias significativas El siguiente interés, entonces, radica en determinar que tan diferentes son los conglomerados con relación a estas variables. La prueba Tukey proporciona los siguientes resultados:

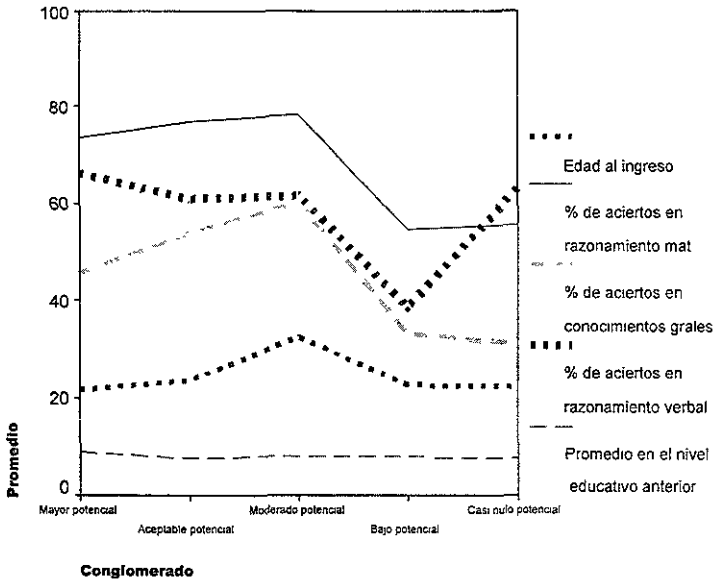
(Ver Anexo 4)

Tabla 6. Resultados de la Prueba Tukey

Variable	Número de grupos que forma la prueba Tukey	Conglomerados agrupados
Edad al ingreso	4	(1,5), (2), (3), (4)
Porcentaje de aciertos en razonamiento matemático	3	(1), (2,3),(5,4)
Porcentaje de aciertos en conocimientos específicos:	4	(1),(2),(3),(4,5)
Porcentaje aciertos en razonamiento verbal	3	(1),(2,3,4),(5)
Promedio en el nivel educativo anterior	4	(1),(2),(3),(4,5)

Para la variable edad al ingreso, porcentaje de respuestas correctas en conocimientos específicos y promedio en el nivel educativo anterior la prueba Tukey forma cuatro grupos y, para las variables porcentaje de respuestas en razonamiento verbal y porcentaje de respuestas en razonamiento matemático, la prueba Tukey determina tres grupos. La gráfica 1 que a continuación se presenta, describe el valor promedio que toma cada una de las variables (no estandarizadas) en cada uno de los conglomerados.

Gráfica 1. Valor promedio de las variables: edad al ingreso, promedio en el nivel educativo anterior y porcentaje de aciertos en el examen de admisión en razonamiento matemático, conocimientos generales y razonamiento verbal en cada uno de los 5 conglomerados.



En la gráfica 1 se observa que el valor medio de la variable promedio en el nivel educativo anterior casi no varía entre los cinco conglomerados, sin embargo, cabe recordar que el valor máximo de esta variable es 10 y el mínimo 6 por lo que, a pesar de que en la gráfica no se observan grandes diferencias entre conglomerados como para las otras variables, no significa que la media del promedio obtenido en la educación media superior no sea diferente en los cinco conglomerados. Así mismo, el valor medio de la edad al ingreso es muy parecida en los dos primeros y en los dos últimos conglomerados. La prueba Tukey presenta para ambas variables cuatro grupos. La siguiente tabla 7 presenta la prueba Tukey para la variable promedio en el nivel educativo anterior y la tabla 8 para la variable edad al ingreso

Tabla 7. Prueba Tukey para la variable promedio en el nivel educativo anterior.

	N	Subset for alpha = .05			
Conglomerados en función de su desempeño académico		1	2	3	4
Aceptable potencial de éxito	210	7 5026			
Casi nulo potencial de éxito	156		7.6765		
Bajo potencial de éxito	172		7.8068		
Moderado potencial de éxito	70			8 0990	
Mayor potencial de éxito	177				8.8508
Mayor potencial de éxito	177				8.8508

Means for groups in homogeneous subsets are displayed.
a Uses Harmonic Mean Sample Size = 135 422

b The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Tabla 8. Prueba Tukey para la variable edad al ingreso.

	N	Subset for alpha = .05			
Conglomerados en función de su desempeño académico		1	2	3	4
Mayor potencial de éxito	177	21 55			
Casi nulo potencial de éxito	156	22 14	22 14		
Bajo potencial de éxito	172		22 58		
Aceptable potencial de éxito	210			23 52	
Moderado potencial de éxito	70				32 40

Means for groups in homogeneous subsets are displayed

a Uses Harmonic Mean Sample Size = 135.422.

b The group sizes are unequal. The harmonic mean of the group sizes is used Type I error levels are not guaranteed

En la tabla 7 se observa que en relación con la variable promedio en el nivel educativo anterior, los conglomerados "casi nulo potencial de éxito" y "bajo potencial de éxito" forman un grupo. El resto de los conglomerados no son agrupados en relación con esta variable. Resalta el hecho que el menor valor promedio de esta variable se observa en el conglomerado "aceptable potencial de éxito". En la tabla 8, para la variable edad al ingreso, se observa que los conglomerados "mayor potencial de éxito" y "casi nulo potencial de éxito" se agrupan. Por su lado, el valor promedio de la edad al ingreso del conglomerado "bajo potencial de éxito" también es muy parecido al de estos dos conglomerados. El valor promedio mayor de la variable edad al ingreso se presenta en el conglomerado "moderado potencial de éxito" (32 40) y el valor promedio menor (21.55) en el conglomerado de "mayor potencial de éxito"

Por otro lado, para las variables relativas al desempeño del alumno en el examen de admisión se observan en la gráfica diferencias muy claras entre conglomerados El porcentaje promedio de aciertos en razonamiento matemático es muy parecido en los conglomerados de "aceptable potencial de éxito" y "moderado potencial de éxito", 76 78% y 78.32% respectivamente. En el caso del conglomerado "mayor potencial de éxito", el porcentaje promedio de aciertos en esta área del examen es del 73.65%. Para los dos últimos conglomerados este porcentaje decrece considerablemente, siendo 55.61% para el caso del conglomerado "bajo potencial de éxito" y de 54 77% para el de "casi nulo potencial de éxito".

El porcentaje promedio de aciertos en conocimientos generales es muy parecido para los conglomerados "casi nulo potencial de éxito" y "bajo potencial de éxito" (30.96% y 33 16%) El porcentaje promedio de aciertos más alto se observa en el conglomerado "aceptable potencial de éxito" (60.19%).

En relación con el área de razonamiento verbal del examen de admisión, se tiene que el valor promedio más alto se registra en el conglomerado "mayor potencial de éxito" (66 17%) siguiéndole el de "bajo potencial de éxito" (63.2%). El porcentaje promedio de aciertos más bajo se presenta en el conglomerado de "bajo potencial de éxito". Estudios anteriores realizados en la universidad muestran que el área de razonamiento verbal es el área del examen de admisión

menos predictiva del desempeño académico de los estudiantes En la siguiente tabla se resumen las características de los estudiantes que integran cada conglomerado:

Tabla 9. Características de los conglomerados

	Conglomerado				
	Mayor potencial de éxito	Aceptable potencial de éxito	Moderado potencial de éxito	Bajo potencial de éxito	Casi nulo potencial de éxito
Edad al ingreso					
Promedio	21.56	23.52	32.39	22.58	22.14
De 17 a 21 años	53.7%	21%	----	39.5%	47.4%
De 22 a 26 años:	44.6%	63.8%	7.1%	49.4%	44.2%
De 27 a31 años.	1.7%	15.2%	37.1%	11%	8.3%
De 32 a 36 años:	---	---	40%	---	---
Más de 36 años:	----	----	15.7%	----	----
Porcentaje de aciertos en conocimientos generales					
Promedio.	45.52%	53.87%	60.20%	33.16%	30.96%
Hasta 25%:	.6%	---	----	17.4%	23.7%
Más de 25% y hasta 50%	70.1%	42.4%	25.7%	79.1%	75.6%
Más de 50% y hasta 75%:	28.2%	52.9%	55.7%	3.5%	.6%
Más de 75%.	1.1%	4.8%	18.6%	----	----
Porcentaje de aciertos en razonamiento matemático.					
Promedio.	73.65%	76.78%	78.32%	54.78%	55.62%
Hasta 25%.	----	----	----	----	----
Más de 25% y hasta 50%	1.1%	1.0%	1.4%	34.9%	27.6%
Más de 50% y hasta 75%.	56.5%	41%	32.9%	62.8%	72.4%
Más de 75%	42.4%	58.1%	65.7%	2.3%	----
Porcentaje de aciertos en razonamiento verbal					
Promedio	66.17%	60.89%	61.62%	38.56%	63.20%
Hasta 25%	---	----	1.4%	4.1%	----
Más de 25% y hasta 50%:	9%	22.4%	10%	93.6%	----
Más de 50% y hasta 75%	69.5%	67.6%	75.7%	2.3%	92.9%
Más de 75%.	21.5%	10%	12.9%	-----	7.1%

Tabla 9. Características de los conglomerados

	Conglomerado				
	Mayor potencial de éxito	Aceptable potencial de éxito	Moderado potencial de éxito	Bajo potencial de éxito	Casi nulo potencial de éxito
Promedio en el nivel educativo anterior					
Promedio.	**8.85**	**7.50**	**8.10**	**7.81**	**7.68**
De 6 hasta 7.	-----	13.8%	8.6%	6.4%	7 7%
Más de 7 hasta 8:	1.7%	76 7%	41.4%	61.6%	70 5%
Más de 8 hasta 9:	61.6%	9.5%	38 6%	29 7%	20.5%
Más de 9:	36.7%	----	11.4%	2.3%	1 3%

En el conglomerado "mayor potencial de éxito", el 53.7% tenían al ingreso entre 17 y 21 años y el 44.6% entre 22 y 26 años; en este conglomerado se encuentran la proporción más alta de estudiantes jóvenes que ingresaron a la universidad. En relación con el examen de admisión, el porcentaje de aciertos en las distintas áreas varía. En el área de conocimientos generales el 70.1% tuvo entre más del 25% y hasta el 50% de aciertos, en el área de razonamiento matemático el 56.5% obtuvo entre más de 50% y hasta 75% de aciertos y el 42.4% más del 75% de aciertos. En el área de razonamiento verbal, la mayoría de los estudiantes (69.5%) obtuvieron entre más del 50% y hasta el 75% de aciertos. Con relación al promedio en la nivel educativo anterior, la mayor parte, 61.6%, obtuvo entre 8 y 9 aunque no deja de ser importante el grupo de alumnos que tenían como promedio en la educación media superior más de 9 (36.7%).

En el conglomerado "aceptable potencial de éxito", el 63.8% tenían al ingreso entre 22 y 26 años. En el área de conocimientos generales del examen de admisión, el 52.9% obtuvo entre más del 50% y hasta el 75% de aciertos y el 42 4% más del 25% y hasta 50% de aciertos El porcentaje de aciertos en el área de razonamiento matemático fue de más del 75% para el 58 1% y de más de 50% y hasta 75% para el 41% de los estudiantes. En razonamiento verbal el 67.6% obtuvo entre más del 50% y hasta 75% de aciertos. En relación al promedio en el nivel educativo anterior, el 76.7% . obtuvo entre 7 y 8.

En el conglomerado "moderado potencial de éxito" no hay alumnos cuya edad al ingreso oscilaba entre 17 y 21 años, el 40% de los integrantes de este conglomerados tenían al ingreso entre 32 y 36 años y el 27.1% de 27 a 31 años. Este es el conglomerado donde se registran las edades más altas al ingreso En relación a las áreas del examen de admisión, más de la mitad (55 7%) obtuvieron más de 50% y hasta 75% de aciertos en el área de conocimientos generales; en el área de razonamiento matemático el 65.7% de los integrantes de este conglomerado

obtuvieron más del 75% de aciertos y en el área de razonamiento verbal, el 75.7% obtuvo entre más del 50% y hasta el 75% de aciertos En lo que se refiere al promedio en la educación media superior, el 41.4% obtuvieron un promedio entre 7 y 8 y un 38 6% entre 8 y 9.

En el conglomerado "bajo potencial de éxito", el 49.4% de los estudiantes ingresaron entre 22 y 26 años y el 39.5% entre 17 y 21 años. En el área de conocimientos generales del examen de admisión, la mayoría obtuvo entre 25% y hasta el 50% de aciertos, en el área de razonamiento matemático se observa que el 62.8% obtuvo entre más de 50% y hasta 75% de aciertos y en el área de razonamiento verbal el 93.6% obtuvo entre más de 25% y hasta 50% de aciertos El promedio del nivel educativo anterior de los estudiantes que integran este conglomerado es de más de 7 y hasta 8 para el 61 6%.

En el conglomerado "casi nulo potencial de éxito", el 47.4% tenía al ingreso entre 17 y 21 años y el 44.2% entre 22 y 26 años. En relación con las áreas del examen de admisión, el 75.6% tuvo entre más del 25% y hasta el 50% de aciertos en el área de conocimientos generales, el 72.4% obtuvo entre más del 50% y hasta el 75% en el área de razonamiento matemático y el 92.9% obtuvo entre más del 50% y hasta el 75% en el área de razonamiento verbal. El promedio *del nivel educativo anterior oscila entre más de 7 y hasta 8 para el 70.5%*

En general se observa que, a pesar de que los estudiantes más jóvenes se encuentran clasificados en el conglomerado de "mayor potencial de éxito", la diferencia de edades de estos estudiantes y los que integran los conglomerados "bajo potencial de éxito" y "casi nulo potencial de éxito" es muy ligera. En el conglomerado de "moderado potencial de éxito" se encuentran los estudiantes más grandes Un alumno joven puede estar tanto en el conglomerado de "mayor potencial de éxito" como en el de "casi nulo potencial de éxito". Claramente se observa que en el conglomerado de "aceptable potencial de éxito" predominan los estudiantes con edades más maduras que en los otros conglomerados.

En razonamiento matemático se observa que para los tres primeros conglomerados el porcentaje de aciertos mayor de 75% en esta área del examen de admisión es casi la mitad o más de la mitad. Para los dos últimos conglomerados el porcentaje de estudiantes que obtuvieron más del 75% es casi nulo. La variable porcentaje de aciertos en el examen de admisión en razonamiento matemático alcanza valores promedio más altos en los cuatro primeros conglomerados en comparación con las otras dos variables relativas al examen de admisión No hay estudiantes en ninguno de los cinco conglomerados cuyo porcentaje de aciertos en el área de razonamiento matemático del examen de admisión sea menor del 25%. Esto verifica el supuesto de que los alumnos que ingresan a ciencias básicas e ingeniería deben tener especiales aptitudes en el área de matemáticas

En conocimientos generales, se observó que el conglomerado cuyos estudiantes tuvieron mejor desempeño en esta área fueron los del conglomerado de "moderado potencial de éxito", que a su vez eran los que tenían mayor edad al ingreso. Los resultados en esta área también son mejores para los dos primeros conglomerados que para los dos últimos como sucede en el caso de aciertos en razonamiento matemático.

En razonamiento verbal resulta difícil establecer una tendencia dado que en los conglomerados de "mayor potencial de éxito", "aceptable potencial de éxito", "moderado potencial de éxito" y casi nulo potencia de éxito se observan resultados muy parecidos.

En el promedio del nivel educativo anterior , los integrantes del conglomerado "mayor potencial de éxito" alcanzaron promedios más altos que los estudiantes de los otros cinco conglomerados. El conglomerado cuyos estudiantes tienen promedio más bajo es el de aceptable potencia de éxito. El conglomerado "moderado potencial de éxito" donde se encuentran los estudiantes cuya edad al ingreso era más alta y cuyo porcentaje de aciertos en el área de conocimientos generales era mayor que para los estudiantes de los otros conglomerados, la media del promedio del nivel educativo anterior es la segunda más alta después de la del conglomerado "mayor potencial de éxito". Los últimos dos conglomerados tienen estudiantes cuyos promedios, en general, son similares.

Cabe señalar que se realizó la prueba no paramétrica Ji-Cuadrada de independencia para probar si existía relación entre cada una de estas variables y la pertenencia a los conglomerados En todos los casos se rechazo la hipótesis nula de independencia (Ver Anexo 4)

Como ya se menciono, el método no jerárquico de k medias fue utilizado para la formación de los conglomerados. Al ser la distancia euclidiana la medida de similaridad no fue posible incluir las variables género y escuela de procedencia. Las siguientes tablas 10 y 11 describen dichas variables para cada conglomerado y las tablas 10.1 y 11.1 las pruebas no paramétricas Ji-cuadrada de independencia respectivamente.

Tabla 10. Género por conglomerado

		Género		Total
		Masculino	Femenino	
Conglomerados en función de su desempeño académico	Mayor potencial de éxito	125 19 7%	52 34 7%	177 22 5%
	Aceptable potencial de éxito	185 29.1%	25 16.7%	210 26 8%
	Moderado potencial de éxito	69 10 9%	1 7%	70 8 9%
	Bajo potencial de éxito	136 21.4%	36 24%	172 21 9%
	Casi nulo potencial de éxito	120 18 9%	36 24%	156 19 9%
Total		635 100%	150 100%	785 100 0%

Tabla 10.1. Prueba Ji-Cuadrada de independencia para género y conglomerado.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	35 243[a]	4	.000
Likelihood Ratio	42 661	4	000
Linear-by-Linear Association	.124	1	725
N of Valid Cases	785		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.38.

En la tabla 10 se observa que aproximadamente la mitad de los estudiantes hombres y la mitad de las mujeres se encuentran en los dos primeros conglomerados Específicamente, el 24.7% de las estudiantes mujeres y el 19.7% de los estudiantes hombres se encuentran en el conglomerado "mayor potencial de éxito" y en el conglomerado aceptable potencial de éxito se encuentra el 29.1% de los hombres y el 16.7% de las mujeres.

En el conglomerado de "moderado potencial de éxito" se observa que solo se encuentra el .7% de las mujeres (una mujer), mientras que el 10.9% de los estudiantes hombres se encuentra clasificado en este grupo. Recordemos que en este conglomerado se encuentran los estudiantes que en su mayoría tenían edad al ingreso de 27 años en adelante.

En el caso de las mujeres se observa que el porcentaje que se encuentra en los dos primeros conglomerados (51.4%) es ligeramente mayor que el porcentaje que se encuentran en los dos últimos conglomerados (48%). Para el caso de los hombres esta relación se mantiene , el

porcentaje que se encuentra en los dos primeros conglomerados es del 48 8% y el porcentaje que se encuentran clasificados en los dos últimos conglomerados es del 40 3%

La prueba no paramétrica Ji-cuadrada de independencia rechaza la hipótesis nula de independencia entre género y pertenencia a los conglomerados

Tabla 11. Escuela de procedencia por conglomerado.

		Escuela de procedencia				Total
		Colegio de Bachilleres	Incorporada a la UNAM	Incorporada a la SEP	Otra[1]	
Conglomerados en función de su desempeño académico	Mayor potencial de éxito	29 11.6%	30 27 5%	77 26.3%	41 31 1%	177 22 5%
	Aceptable potencial de éxito	66 26 3%	37 33.9%	76 25 9%	31 23 5%	210 26.8%
	Moderado potencial de éxito	26 10 4%	7 6.4%	24 8 2%	13 9 8%	70 8 9%
	Bajo potencial de éxito	74 29 5%	14 12 8%	59 20 1%	25 18.9%	172 21.9%
	Casi nulo potencial de éxito	56 22 3%	21 19 3%	57 19 5%	22 16 7%	156 19 9%
Total		251 100%	109 100%	293 100%	132 100%	785 100 0%

Tabla 11.1. Prueba Ji-Cuadrada de independencia para escuela de procedencia en el nivel educativo anterior y conglomerado.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	38 167[a]	12	.000
Likelihood Ratio	40 532	12	000
Linear-by-Linear Association	14 912	1	000
N of Valid Cases	785		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 9 72

En la tabla 11 se observa que para los estudiantes que egresaron del Colegio de Bachilleres no hay una tendencia específica ya que el 26.3% d ellos se encuentra en el conglomerado de "aceptable potencial de éxito", el 29.5% en el de "bajo potencial de éxito" y el

[1] IPN CECyT, Normal primaria, Universidad Estatal, Incorporada a Universidad Estatal y cualquier otro tipo de Institución que avale certificados de Educacion Media Superior

22.3% en el de "casi nulo potencial de éxito". Su presencia en el conglomerado de "mayor potencial de éxito" es de apenas 11.6%

Para los estudiantes que finalizaron su educación media superior en escuelas incorporadas a la UNAM, se observa que más de la mitad de los estudiantes, 51.4%, se encuentran en los conglomerados "mayor potencial de éxito" y "aceptable potencial de éxito" (27.5% y 33.9%). Sin embargo, casi una quinta parte de ellos se encuentra en el conglomerado de "casi nulo potencial de éxito" (19 2%).

El 52.2% de los alumnos egresados de escuelas incorporadas a la SEP se encuentran clasificados en los conglomerados de "mayor potencial de éxito" y "aceptable potencial de éxito" Sin embargo la proporción de ellos que se encuentran en los últimos dos conglomerados no deja de ser importante, 20.1% en el conglomerado "bajo potencial de éxito" y 19 5% en el conglomerado "casi nulo potencial de éxito"

De los estudiantes que provienen de escuelas clasificadas como Otras, se observa que el 31.1% se encuentra en el conglomerado de "mayor potencial de éxito" y el 23.5% en el de "moderado potencial de éxito". Sin embargo, el 18.9% se encuentra en el conglomerado de "bajo potencial de éxito" y el 16 7% en el de "casi nulo potencial de éxito".

La prueba no paramétrica Ji-cuadrada de independencia rechaza la hipótesis nula de independencia entre escuela de procedencia del nivel educativo anterior y la pertenencia a los conglomerados

Recordemos que con el análisis de conglomerados se formaron 5 grupos con características homogéneas. Hasta aquí hemos descrito cada uno de los conglomerados en función de las características de los estudiantes que los componen. La siguiente etapa del análisis de datos consiste en estimar la función de sobrevivencia para cada uno de estos conglomerados y compararlas con el objeto de establecer si hay diferencias significativas entre ellas.

3.4. Análisis de Sobrevivencia

3.4.1. Introducción.

La principal medida en el análisis de sobrevivencia es el tiempo transcurrido hasta un evento crítico. La variable tiempo debe tener al menos propiedades de escala de intervalo, lo cual puede sugerir métodos tales como regresión y análisis de varianza. Sin embargo, los estudios de sobrevivencia generalmente incluyen datos censurados que son datos para los cuales no se conoce el tiempo para el evento crítico. El análisis de sobrevivencia, a diferencia del análisis de regresión y el análisis de varianza. consideran estos casos.

El tiempo promedio de sobrevivencia y la mediana del tiempo de sobrevivencia es estimado utilizando la función acumulada de sobrevivencia la cual ajusta los datos censurados. Un resumen en tabla o gráfico que contenga información acerca del a función de sobrevivencia a través del tiempo representa una herramienta útil en estudios longitudinales donde se busca estimar la probabilidad de que ocurra un evento crítico en un determinado periodo de tiempo.

La curva de la función acumulada de sobrevivencia es una curva empírica ajustada para cada evento crítico. Esta aproximación no asume distribuciones por tanto es no paramétrica y se le conoce como la aproximación producto-límite de Kaplan y Meier. Otra función en el análisis de sobrevivencia es la función de riesgo, que mide la tasa de ocurrencia por unidad de tiempo de un evento crítico en un instante . La función de sobrevivencia está relacionada con la función de azar

Una característica distinguible del Análisis de sobrevivencia es que utiliza casos censurados que son casos en que el evento de interés no ha ocurrido al final del estudio o en un tiempo de análisis determinado y por tanto se desconoce el tiempo exacto de sobrevivencia. Otras técnicas estadísticas utilizan estos datos como datos perdidos ya que son observaciones para los cuales el valor de interés se desconoce. En este método los casos censurados son utilizados para calcular la función de sobrevivencia.

Cuando se trabaja con un grupo, es de interés examinar la curva acumulada de sobrevivencia que muestra la estimación de la probabilidad de sobrevivencia más allá del final de cada periodo de tiempo. Más aún, se obtiene la media o mediana de los tiempos de sobrevivencia con sus respectivos errores estándares.

Cuando se trabaja con más de un grupo, es de interés la comparación de la función acumulada de sobrevivencia de cada grupo. Además de la media y mediana del tiempo de

sobrevivencia y pruebas de significación de diferencias entre las funciones de sobrevivencia de los diferentes grupos.

Finalmente, si se tienen datos medidos en variables categóricas o variables de intervalo, el Análisis de sobrevivencia resulta predictivo a través de la regresión de Cox

3.4.2. Definición del problema

Evento crítico: no - retención

El evento crítico no – retención es observado en aquellos alumnos que durante más de seis periodos consecutivos no inscribieron asignaturas o que su desempeño acumulado al doceavo trimestre no les permitiría, ni en los dieciocho trimestres restantes que la universidad establece como periodo máximo de permanencia en la universidad, acreditar todas las asignaturas (UEA´s[1]) establecidas en el plan de estudios de su carrera respectiva

Cabe mencionar que si se quisiera establecer como evento crítico a la deserción, la información que se dispone no resulta suficiente para establecer que un alumno ha desertado Principalmente esto se debe a que el Reglamento de Estudios Superiores de la universidad establece como plazo máximo para la aprobación del 100% de créditos correspondientes a las Unidades de Enseñanza Aprendizaje (UEA's) respectivas de cada una de las carreras 10 años que equivalen a 30 trimestres Por tanto, si no se cuenta con la información de 10 años de una cohorte determinada no es posible determinar con certeza si un alumno es desertor.

Tiempo de sobrevivencia:

Tiempo en el que se ocurre el estado crítico no – retención o tiempo censurado.

Tiempo de observación:

12 trimestres

Observaciones censuradas:

Las observaciones censuradas son aquellas que al tiempo 12 no han manifestado el evento crítico no - retención. De hecho, las observaciones censuradas son estudiantes rezagados que de alguna manera siguen cursando materias en la universidad.

Según la clasificación del capítulo 2 (Apartado 2 4.1.2) son del Tipo I, dado que el tiempo de observación es de 12 trimestres y a este tiempo el evento crítico no - retención no ha ocurrido

Estado:

El estado de una observación, es decir, el estado de un alumno puede ser

- Alumno que ha experimentado el evento crítico no - retención

- Alumno en que al tiempo de observación, doceavo trimestre, no se ha presentado el evento crítico no – retención. Este tipo de alumnos son observaciones censuradas. El tiempo de sobrevivencia es 12^{+} trimestres. Es importante resaltar que si un alumno es observación censurada significa que aún permanece con carácter de alumno en la universidad y que al menos no ha abandonado sus estudios en 6 trimestres consecutivos y que su desempeño académico le permitiría alcanzar el 100% de créditos antes de que el plazo máximo de treinta trimestres tome vigencia y por tanto aún puede tener éxito en sus estudios universitarios.

3.4.3. Obtención de los tiempo de sobrevivencia

Se tiene una población de 785 alumnos que ingresaron en los periodos primavera y otoño de 1995 a las 9 diferentes carreras que imparte la división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana – Iztapalapa.

La obtención de los tiempos de sobrevivencia se llevó a cabo principalmente con dos criterios:

1 Unidades de Enseñanza Aprendizaje inscritas:

2. Desempeño académico:

1. Unidades de Enseñanza Aprendizaje inscritas.

I. La Universidad Autónoma Metropolitana denomina Unidades de Enseñanza Aprendizaje a cada una de las asignaturas de las carreras que imparte. De los 785 estudiantes, 140 de ellos no inscribieron ninguna materia desde el segundo trimestre hasta el doceavo El tiempo de sobrevivencia para estos casos corresponde al segundo trimestre ya que el estado crítico no-retención ocurre en este periodo.

[1] UEA: Unidad de Enseñanza Aprendizaje. Este concepto es utilizado en la Universidad Autónoma Metropolitana para

El porcentaje de estudiantes que no inscribieron asignaturas desde el primer trimestre es casi nulo (.38%) Esto se debe, primero a que la universidad asigna un bloque de materias a todos los alumnos de primer ingreso y segundo, a que la universidad permite al alumno, en un periodo determinado posterior a la inscripción, dar de baja las materias que ya sea el haya inscrito o se le hayan asignado por primer ingreso En el caso de primer ingreso, el alumno respondiendo a un proceso de adaptación y conocimiento de la universidad, en general no utiliza este recurso Para estos casos el tiempo de sobrevivencia es el primer trimestre pero por ahora solo nos ocuparemos de los estudiantes cuyo abandono fue a partir del segundo trimestre

Cabe señalar que 121 de los 140 (86.42%) alumnos que no inscribieron asignaturas a partir del segundo trimestre, tuvieron un desempeño académico nulo durante los 12 trimestres, es decir, no acreditaron ninguna materia durante el periodo señalado como regular por la universidad.

Tabla 1. Distribución de alumnos por carrera que no inscribieron asignaturas a partir del segundo trimestre

Carrera	Número de estudiantes	Porcentaje[1]	Porcentaje con relación al nuevo ingreso[2]
Ingeniería Biomédica	10	7 1	12.0%
Ingeniería Hidrológica	7	5 0	26 9%
Ingeniería Química	17	12 1	19 5%
Ingeniería en energía	18	12 9	22 0%
Física	21	15 0	30 4%
Ingeniería Electrónica	13	9.3	9 6%
Matemáticas	25	17.9	18 5%
Química	17	12 1	39 5%
Computación	12	8.6	9 6%
Total	140	100 0	

1. Porcentaje de alumnos que no inscribieron asignaturas a partir del segundo trimestre.
2. Indica el porcentaje de los de nuevo ingreso de cada carrera que no inscribieron asignaturas a partir del segundo trimestre.

Se observa que en la licenciatura en Química el 39.5% de los alumnos de nuevo ingreso no inscribieron ninguna asignatura a partir del segundo trimestre. A su vez, los resultados son también alarmantes en la licenciatura en Física e Ingeniería Hidrológica, 30.4% y 26.9% respectivamente

referirse a cada una de las asignaturas de las carreras que imparte.

II. Por reglamento en la Universidad Autónoma Metropolitana, si un alumno abandona la universidad por más de 6 meses pierde su carácter de alumno y para recuperarlo debe presentar un examen de las asignaturas (unidades de enseñanza aprendizaje) acreditadas antes de este abandono. El porcentaje de alumnos que utilizan este proceso para recuperar su carácter de alumno es casi nulo. Por tanto, se considerará que los alumnos que no hayan inscrito materias por más de seis periodos consecutivos, el tiempo al estado crítico no-retención es el último periodo antes de este evento. Así si el estudiante no inscribió materias los 6 primeros trimestres, el tiempo de sobrevivencia es 1, si el estudiante no inscribió créditos a partir del trimestre 2 y durante 6 trimestres consecutivos el tiempo de sobrevivencia es 2 y así sucesivamente Por tanto, se tienen 6 diferentes tiempos de sobrevivencia.

Tabla 2. Tiempos de sobrevivencia para alumnos que durante 6 o más periodos no inscribieron ninguna materia.

Tiempo de sobrevivencia	Número de estudiantes	Porcentaje
1	3	1.3%
2	7	3.0%
3	63	26.9%
4	57	24.4%
5	36	15.4%
6	40	17 1%
7	28	12 0%
Total	234	100.0

2. Por desempeño académico

I. La universidad establece como plazo máximo para terminar los estudios de licenciatura 30 trimestres. El número de créditos mínimos para considerar al alumno sobreviviente en cada uno de los trimestres es la división del número total de créditos de la carrera entre 30 y multiplicado por el trimestre. Por ejemplo, la carrera de Ingeniería Biomédica es de 539 créditos, un alumno es sobreviviente del trimestre 1 si el número de créditos aprobados en el trimestre 1 es mayor que el resultado de la división de 539 entre 30. Así mismo, el alumno es sobreviviente en el periodo 2 si el número de créditos aprobados hasta el periodo 2 es mayor o igual que 530 entre 30 y por 2.

Éste pudiera parecer un criterio simple para establecer los tiempos de sobrevivencia de los estudiantes. Sin embargo es una realidad que el alumno pudo no haber acreditado el mínimo de créditos con relación al tiempo máximo en uno de los trimestres pero en el o los siguientes trimestres posteriores pudo haber alcanzado este parámetro. Esto sugiere que para determinar el tiempo de sobrevivencia es necesario detectar a los alumnos que al trimestre 12 no cumplen con este parámetro, es decir, que el número de créditos aprobados hasta el trimestre 12 sea menor que el número de total de créditos de la carrera entre 30 y multiplicado por 12. El tiempo de sobrevivencia de dichos alumnos será el último trimestre en el que inscribieron materias más uno. Por ejemplo, un alumno que no inscribe unidades de enseñanza aprendizaje a excepción del primer y quinto trimestre registra un tiempo de sobrevivencia de 6 dado que el evento crítico de no-retención se presenta en el trimestre 6 ya que en el trimestre 5 el alumno todavía se inscribió.

La distribución de los tiempos de sobrevivencia según esta consideración se resume en la siguiente tabla 3.

Tabla 3. Distribución de los tiempos de sobrevivencia según primer criterio[1] de desempeño académico.

Tiempo de sobrevivencia	Número de estudiantes	Porcentaje
8	21	8 5
9	11	4.5
10	16	6.5
11	40	16 3
12	158	64.2
Total	246	100 0

1.El primer criterio consiste en determinar el último trimestre en el que inscribieron asignaturas alumnos cuyo avance de créditos al doceavo trimestre es menor que el (número total de créditos de la carrera / 30)*12 El tiempo de sobrevivencia es dicho trimestre más uno

II. El desempeño académico de los 165 estudiantes restantes se ilustra en la tabla 4

Tabla 4. Desempeño académico al trimestre 12 de estudiantes cuyo avance de créditos es mayor o igual al (número total de créditos de la carrera / 30)*12 y que no han dejado de inscribir asignaturas por más de 6 trimestres.

	Número de estudiantes	Porcentaje
Del 25% hasta el 50%	46	27 9
Del 50% hasta el 75%	91	55.2
Más del 75%	28	17.0
Total	165	100.0

Se observa que más de la mitad registran un desempeño académico entre el 50% y el 75%. En general, los 165 estudiantes aunque no cumplen con el avance de créditos establecido por la universidad, se caracterizan por la constancia durante sus estudios

De los 165 estudiantes, 129 inscribieron materias de manera constante durante los 11 primeros periodos A estos alumnos, se les asigna el tiempo sobrevivencia 12, ya que el estado crítico no - retención no es observado al menos hasta este trimestre. Cabe señalar que 24 de estos 129 estudiantes sólo han cubierto entre el 25% y el 50% de los créditos.

Los 36 alumnos restantes son estudiantes que en algún trimestre no inscribieron materias. A pesar de este hecho, en general, presentan constancia ya que no se inscriben en un determinado trimestre pero antes de seis trimestres ya se volvieron a inscribir. El tiempo de sobrevivencia en estos casos también será de 12 ya que tampoco se observa el estado crítico de no - retención.

Es notorio que ninguno de estos 160 estudiantes inscribió materias en el doceavo trimestre.

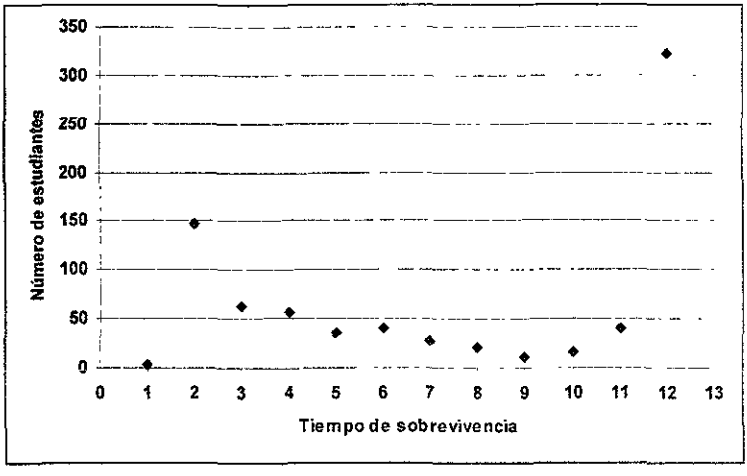
La distribución de los tiempos de sobrevivencia al estado crítico no-retención para los 785 casos se presenta en la siguiente tabla 5

Tabla5. Tiempos de sobrevivencia de los 785 estudiantes

Tiempo de sobrevivencia	Frecuencia	Porcentaje	Porcentaje acumulado
1	3	.4	.4
2	147	18.7	19 1
3	63	8 0	27 1
4	57	7.3	34.4
5	36	4.6	39.0
6	40	5 1	44.1
7	28	3.6	47 6
8	21	2.7	50 3
8	21	2 7	50 3
9	11	1 4	51.7
9	11	1.4	51 7
10	16	2.0	53.8
10	16	2.0	53.8
11	40	5 1	58 9

12	323	41 1	100 0
Total	785	100 0	

Gráfica 1. Tiempos de sobrevivencia.



En la tabla 5 se observa que el 41.1% son observaciones censuradas, es decir, al trimestre 12 no se les observo el evento crítico de no-retención. El 58.9% son estudiantes que no fueron retenidos por la universidad. Ambas cifras reportan el alto índice de rezago y abandono en la División de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana, Unidad Iztapalapa. Cabe resaltar que de los 785 casos no hay ningún alumno cuyo desempeño académico al doceavo trimestre haya sido del 100%.

3.4.4. Método no paramétrico producto-límite de Kaplan y Meier

El procedimiento de Kaplan Meier es un método no paramétrico que estima la función de sobrevivencia cuando hay observaciones censuradas, es decir, observaciones para las cuales no se conoce el tiempo exacto de sobrevivencia. La función de sobrevivencia es decreciente y su valor máximo es de uno al tiempo de inicio. El método de Kaplan Meier puede realizar pruebas de significación comparando grupos que difieren en un solo factor. Además, es posible especificar una segunda variable de grupo. Esto permite llevar a cabo pruebas de comparación entre los niveles de factor en forma independiente para cada nivel de la variable grupo

Dado el carácter comparativo de esta investigación es de interés la obtención de estadísticos que permitan probar diferencias en las distribuciones de sobrevivencia entre los conglomerados ya que con ello es posible determinar si un grupo tiene función de sobrevivencia diferente a otro y de este modo encontrar quienes son los estudiantes que son más propensos a que les ocurra el evento crítico no - retención. Para esto se realizarán las pruebas Log Rank, Breslow y Tarone-Ware. Cada una de estas pruebas estadísticas, está basada en la comparación del número de eventos críticos observados y el número de eventos esperados en cada periodo de tiempo. El número de eventos esperados se deriva del número de casos en riesgo y el número para los que el evento crítico ocurrió en determinado periodo de tiempo. Si no hay diferencia entre lo niveles del factor entonces el número de eventos esperado debe ser cercano al número observado para los diferentes niveles del factor. Estas pruebas difieren en el peso asignado a cada uno de los eventos cuando se calcula el estadístico final .La prueba Log Rank asigna el mismo peso a un evento que haya ocurrido antes o después en la escala de tiempo. La prueba Breslow asigna pesos a los eventos con base en los casos que están en riesgo. entonces el número de casos en riesgo disminuye a través del tiempo, de tal forma que los eventos tempranos tienen peso mayor que los eventos posteriores. La prueba Tarone-Ware asigna pesos a los eventos por medio de la raíz cuadrada del número de casos en riesgo y por tanto el peso de los eventos es aún menor que en la prueba de Breslow pero mayor que en la prueba de Log Rank.

Variables utilizadas para el análisis:

Tiempo:

Tiempo al evento crítico no-retención o tiempo censurado.

Estado:

Variable que indica si el evento crítico sucedió o es una observación censurada.

Factor

1. Conglomerado
2. Conglomerado por género y escuela de procedencia
3. Género.
4. Edad.
5. Escuela de Procedencia en el nivel educativo anterior.
6. Promedio obtenido en el nivel educativo anterior.
7. Porcentaje de aciertos en el examen de admisión en las tres diferentes áreas: conocimientos generales, razonamiento matemático y razonamiento verbal.
8. Periodo de ingreso.
9. Carrera

Resultados solicitados al paquete estadístico SPSS (Statistical Programme for Social Sciences)

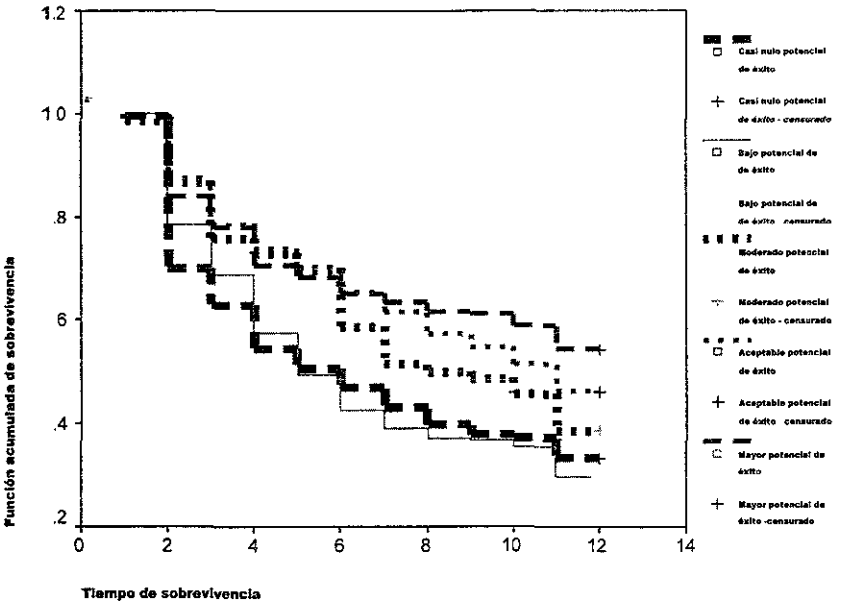
1. Curva o función de sobrevivencia (función acumulada de sobrevivencia)
2. Media y mediana del tiempo de sobrevivencia.
3. Prueba de comparación de funciones de sobrevivencia Log Rank, Tarone Ware y Breslow

En el Anexo 5 se encuentran los reportes completos obtenidos del SPSS.

3.4.4.1. Método no paramétrico producto-límite de Kaplan y Meier por Conglomerado.
(ver anexo 5.1)

La gráfica 1 presenta la función acumulada de sobrevivencia de cada uno de los conglomerados. En general se observa que la probabilidad de sobrevivir al evento crítico no - retención es mayor en el conglomerado con "mayor potencial de éxito" siguiéndole el de "aceptable potencial de éxito" y después el de "moderado potencial de éxito". Para los dos últimos conglomerados se observa que a partir del quinto trimestre la probabilidad de sobrevivencia al evento crítico de los estudiantes que integran el conglomerado de "bajo potencial de éxito" es mayor que la de los estudiantes que integran el conglomerado "casi nulo potencial de éxito"

Gráfica 1. Funciones de sobrevivencia de los conglomerados



Las medianas estimadas del tiempo de sobrevivencia se presentan en la siguiente tabla 1.

Tabla 1. Medianas estimadas del tiempo de sobrevivencia para cada conglomerados.

Conglomerado	Mediana del tiempo de sobrevivencia
1.Mayor potencial de éxito	-
2 Aceptable potencial de éxito	11
3 Moderado potencial de éxito	8
4.Bajo potencial de éxito	5
5.Casi nulo potencial de éxito	6

– significa que al doceavo trimestre el evento crítico no había ocurrido en al menos el 50% de las observaciones

Al trimestre 12, al menos el 50% de los estudiantes que integran el primer conglomerado sobrevivieron al evento crítico no - retención. Por otro lado, la mediana estimada del tiempo de sobrevivencia para el conglomerado "aceptable potencial de éxito" es el onceavo trimestre y para el de "moderado potencial de éxito" es el octavo trimestre, es decir, para el primer caso en el trimestre once al menos al 50% de los estudiantes que integran este conglomerado ya les había ocurrido el evento crítico no - retención y para el caso del conglomerado "moderado potencial de éxito", al

menos el 50% de los estudiantes eran no retenidos en la universidad en el octavo trimestre. La mediana estimada del conglomerado "bajo potencial de éxito" es más alta que la mediana estimada del conglomerado "casi nulo potencial de éxito", trimestre cinco y trimestre seis respectivamente.

El porcentaje de casos censurados es de más del 45% en los dos primeros conglomerados y de 38.6% para el conglomerado "aceptable potencial de éxito". Por su lado el porcentaje de observaciones censuradas en los dos últimos conglomerados es del 29.6% en el conglomerado de "bajo potencial de éxito" y de 33 3% en el conglomerado "casi nulo potencial de éxito". Recordemos que un estudiante se considera observación censurada si al final del estudio, es decir, al doceavo trimestre, no le ha ocurrido el evento crítico no - retención. Por tanto, un estudiante considerado observación censurada es un alumno activo rezagado.

Como se observa en la tabla 2, al nivel de significancia α= 05, las pruebas Log Rank, Breslow y Tarone-Ware establecen diferencias significativas entre las funciones de sobrevivencia de los 5 conglomerados.

Tabla 2. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia.

	Estadístico	gl	Significancia
Log Rank	31.87	4	.0000
Breslow	32.08	4	.0000
Tarone-Ware	32.39	4	.0000

Entonces, el siguiente interés radica en conocer quiénes son diferentes. La siguiente tabla 3 presenta las parejas de conglomerados que en al menos una prueba presentaron diferencias significativas entre sus funciones de sobrevivencia.

Tabla 3. Pruebas Log Rank, Breslow y Tarone Ware conglomerado a conglomerado

Casos en los que la Ho de igualdad de funciones es rechazada.	Log Rank	Breslow	Tarone – Ware
(4,1)	☑	☑	☑
(4,2)	☑	☑	☑
(4,3)	—	☑	☑
(5,1)	☑	☑	☑
(5,2)	☑	☑	☑
(5,3)	----	☑	----

Se observa, como era de esperarse, que la función de sobrevivencia de los conglomerados "bajo potencial de éxito" y "casi nulo potencial de éxito" presentan diferencias significativas con cada uno de los tres primeros conglomerados. Las parejas de conglomerados que no fueron

registradas en la tabla anterior se deben a que la hipótesis nula de igualdad de funciones de sobrevivencia no fue rechazada en ninguna de las tres pruebas.

Dado que en la formación de los conglomerados no se incluyen las variables categóricas género y escuela de procedencia en el nivel educativo anterior, se realizará el análisis de sobrevivencia para los conglomerados diferenciándolos por estas variables

3.4.4.2. Método no paramétrico producto-límite de Kaplan y Meier por conglomerado, género y escuela de procedencia en el nivel educativo anterior.

3.4.4.2.1. Escuela de procedencia: Colegio de Bachilleres

(ver anexo 5.2 1)

Tanto para el caso de estudiantes mujeres como para el caso de estudiantes hombres, las pruebas Log Rank , Breslow y Tarone Ware, al nivel de significancia α=.05, no rechazan la hipótesis nula de igualdad de funciones de sobrevivencia de los conglomerados

Tabla 4. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia de los conglomerados y para estudiantes de género masculino.

	Estadístico	gl	Significancia
Log Rank	8.13	4	.0868
Breslow	6.75	4	.1494
Tarone-Ware	7.51	4	.1112

Tabla 5. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia de los conglomerados y para estudiantes de género femenino.

	Estadístico	gl	Significancia
Log Rank	.75	3	.8611
Breslow	.66	3	.8828
Tarone-Ware	.70	3	.8737

Recordando que los conglomerados fueron formados por estudiantes con características homogéneas y después modelados en función de su desempeño académico al doceavo trimestre, que no haya diferencias significativas entre las funciones de sobrevivencia de los conglomerados sugiere que, independientemente del conglomerado al que pertenezcan, no hay diferencias significativas en la probabilidad de sobrevivencia de los estudiantes hombres que obtuvieron su

certificado de educación media superior en el Colegio Bachilleres El mismo resultado se observa para estudiantes del género femenino egresadas del Colegio de Bachilleres

El porcentaje total de casos censurados para estudiantes de género femenino es 46 7% mientras que para el caso de los hombres es el 35 9%. Además se infiere que, tanto para mujeres como para hombres, el que haber egresado del Colegio de Bachilleres no es un factor que determina su permanencia en la universidad

3.4.4.2.2. Escuela de procedencia: Incorporada a la UNAM

(ver anexo 5 2.2)

Tanto para el caso de estudiantes de género masculino como para el caso de estudiantes de género femenino, las pruebas Log Rank, Breslow y Tarone-Ware, al nivel de significancia α=.05, no rechazan igualdad de funciones de sobrevivencia de los 5 conglomerados Esto sugiere que para ambos género, el ser egresado de escuelas incorporadas a la UNAM no determina la probabilidad de ser retenido en la universidad.

Tabla 6. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia de los conglomerados y para estudiantes de género masculino.

	Estadístico	gl	Significancia
Log Rank	5.66	4	.2258
Breslow	5.88	4	.2086
Tarone-Ware	5.76	4	.2180

Tabla 7. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia de los conglomerados y para estudiantes de género femenino.

	Estadístico	gl	Significancia
Log Rank	3.51	3	.3192
Breslow	4.41	3	.2203
Tarone-Ware	3.97	3	.2644

El porcentaje total de casos censurados en el caso de las mujeres es del 54.6% mientras que para los hombres es del 46%. Cabe mencionar que los estudiantes que registran tiempos censurados, si bien no han acreditado el 100% de las Unidades de Enseñanza Aprendizaje tampoco han sino no - retenidos en la universidad.

3.4.4.2.3. Escuela de procedencia: Incorporada a la SEP

(ver anexo 5.2.3)

Al nivel de significación α=.05, las pruebas estadísticas Log Rank, Breslow y Tarone-Ware establecen diferencias significativas entre las funciones de sobrevivencia de los conglomerados integrados por estudiantes de género masculino y cuya escuela de procedencia en el nivel educativo anterior fue incorporado a la SEP. Esto sugiere que para los hombres, el hecho de haber cursado su educación media superior en escuelas incorporadas a la SEP de alguna manera determina su probabilidad de ser retenido en la universidad.

Tabla 8. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia de los conglomerados y para estudiantes de género masculino.

	Estadístico	gl	Significancia
Log Rank	17.54	4	.0015
Breslow	17.98	4	.0012
Tarone-Ware	18.04	4	.0012

Tabla 9. Pruebas Log Rank, Breslow y Tarone Ware conglomerado a conglomerado de estudiantes de género masculino y de escuela de procedencia en el nivel educativo anterior Escuela incorporada a la SEP.

Casos en los que la Ho es rechazada.	Log Rank	Breslow	Tarone – Ware
(4,1)	☑	☑	☑
(4,2)	☑	☑	☑
(4,3)	—	☑	—
(5,1)	☑	☑	☑
(5,2)	☑	☑	☑
(5,3)	—	☑	—

El interés, entonces, radica en determinar las parejas de conglomerados que son diferentes. En la tabla 9 se observa que la función de sobrevivencia del conglomerado "bajo potencial de éxito" y la del conglomerado "casi nulo potencial de éxito" presenta diferencias significativas con las funciones de sobrevivencia de los tres conglomerados restantes que son precisamente en los que se espera que tengan probabilidades más altas de sobrevivencia dado que en ellos se encuentran estudiantes con desempeño académico mejor.

Algunas parejas de conglomerados no aparecen en la tabla 9 dado que la hipótesis nula de igualdad de funciones de sobrevivencia no fue rechazada en ninguna de las tres pruebas utilizadas

Como se observa en la tabla 10, la mediana del tiempo de sobrevivencia disminuye de acuerdo a la jerarquización de los conglomerados. Recordemos que las pruebas Log Rank, Tarone Ware y Breslow establecieron diferencias significativas entre las funciones de sobrevivencia de los conglomerados "casi nulo potencial de éxito" y "bajo potencial de éxito" con cada uno de los tres conglomerados restantes

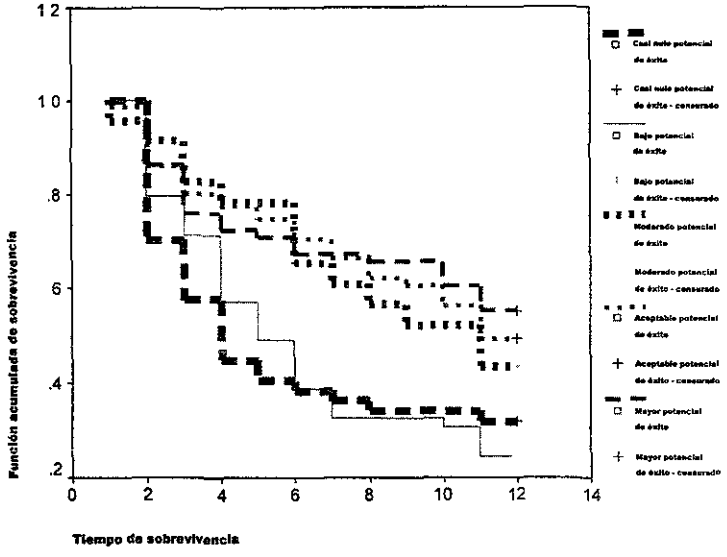
Tabla 10. Medianas del tiempo de sobrevivencia por conglomerado para estudiantes de género masculino egresados de escuelas incorporadas a la SEP:

Conglomerado	Mediana del tiempo de sobrevivencia Género: Masculino
1. Mayor potencial éxito	-
2. Aceptable potencial de éxito	11
3. Moderado potencial de éxito	11
4. Bajo potencial de éxito	5
5. Casi nulo potencial de éxito	4

- significa que al doceavo trimestre el evento crítico no había ocurrido en al menos el 50% de las observaciones

En la gráfica 2, donde se ilustra la función de sobrevivencia por conglomerado en el caso de estudiantes de género masculino, se observa que los conglomerados "mayor potencial de éxito", "aceptable potencial de éxito" y "moderado potencial de éxito" forman un grupo y los últimos dos conglomerados otro grupo. Claramente la probabilidad de sobrevivencia de los integrantes de cualquiera de los tres primeros conglomerados es mayor que en el caso de los estudiantes que integran los conglomerados "bajo potencial de éxito" y "casi nulo potencial de éxito"

Gráfica 2. Funciones de sobrevivencia por conglomerado de estudiantes de género masculino y de escuela de procedencia en el nivel educativo anterior incorporada a la SEP.



Para el caso de estudiantes de género femenino, la igualdad de funciones de sobrevivencia de los conglomerados no se rechaza al nivel de significación α=.05, es decir, en el caso de las mujeres el haber terminado su educación media superior en Instituciones incorporadas a la SEP no es un factor que determina su desempeño académico o su permanencia en la universidad.

Tabla 11. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia de los conglomerados y para estudiantes de género femenino.

	Estadístico	gl	Significancia
Log Rank	4.54	4	.3374
Breslow	3.57	4	.4666
Tarone-Ware	4.04	4	.4009

El porcentaje de casos censurados en las mujeres es del 51.1% y 41.94% en el caso de los hombres

3.4.4.2.4. Escuela de procedencia: Otra[1]

(ver anexo 5 2 4)

Para ambos géneros, al nivel de significancia de α= 05, las prueba Log Rank, Breslow y Tarone-Ware, rechazan la hipótesis nula de igualdad de funciones de sobrevivencia en los 5 conglomerados Esto sugiere que tanto para estudiantes hombres como para estudiantes mujeres, el haber egresado de instituciones educativas clasificadas como Otras no determina su probabilidad de retención en la universidad.

Tabla 12. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia de los conglomerados y para estudiantes de género masculino.

	Estadístico	gl	Significancia
Log Rank	5.55	4	.2351
Breslow	6.66	4	.1550
Tarone-Ware	6.06	4	.1948

Tabla 13. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia de los conglomerados y para estudiantes de género femenino.

	Estadístico	gl	Significancia
Log Rank	3.39	3	.3358
Breslow	3.15	3	.3697
Tarone-Ware	3.30	3	.3480

El porcentaje de observaciones censuradas es de 37.14% para el caso de los estudiantes hombres y del 33.3% para el caso de estudiantes mujeres.

3.4.4.2.5. Alcances

Las funciones de sobrevivencia de los 5 conglomerados, que no consideran el género del estudiante y la escuela de procedencia en el nivel educativo anterior, se distribuyen jerárquicamente según las etiquetas que se les asignaron, a excepción de los últimos dos conglomerado en los que a partir del cuarto trimestre la función de sobrevivencia del conglomerado "casi nulo potencial de éxito" está por encima de la función "bajo potencial de éxito". De este modo, los integrantes del

[1] IPN CECyT, Normal primaria, Universidad Estatal, Incorporada a Universidad Estatal y cualquier otro tipo de Institución que avale certificados de Educación Media Superior.

conglomerado "mayor potencial de éxito" tienen mayor probabilidad de sobrevivir que los estudiantes que integran el conglomerado "moderado potencial de éxito" y a su vez los integrantes de este conglomerado tienen mayor probabilidad de sobrevivir que los estudiantes que componen el conglomerado "moderado potencial de éxito". Este resultado, hasta cierto punto, verifica el supuesto de que a mejor desempeño académico de un estudiante, la probabilidad de ser retenido en la universidad es mayor.

- Al integrar las variables género y escuela en el nivel educativo anterior se concluye lo siguiente

    1. Salvo en el caso de los alumnos que provienen de escuelas clasificadas como Otras, las mujeres tienen mayor probabilidad de sobrevivir al evento crítico no - retención que los hombres.

    2. En ambos géneros, los estudiantes cuya escuela de procedencia en la educación media superior fue incorporada a la UNAM son los que presentan mayor probabilidad de no experimentar el evento crítico no - retención.

    3. La probabilidad de sobrevivencia al evento crítico no - retención solo es determinado por la escuela de procedencia en el caso de los hombres que egresan de escuelas incorporadas a la SEP

3.4.4.3. Método no paramétrico producto-límite de Kaplan y Meier por género.
(ver anexo 5.3)

Al nivel de significancia α=.05, las pruebas Log Rank, Breslow y Tarone-Ware, no rechazan la hipótesis nula de igualdad de funciones de sobrevivencia. Es decir, no hay diferencias significativas entre las funciones de sobrevivencia de los estudiantes hombres y de los estudiantes mujeres.

Tabla 14. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia

	Estadístico	gl	Significancia
Log Rank	3.05	1	.0806
Breslow	2.70	1	.1006
Tarone-Ware	2.94	1	.0863

En el caso de las mujeres el 47% de las observaciones registran tiempos censurados, es decir, estudiantes que al trimestre 12 aún no habían sido no - retenidos en la universidad En este mismo sentido, el 39.69% de los hombres son observaciones censuradas

3.4.4.1. Método no paramétrico producto-límite de Kaplan y Meier por edad al ingreso. (ver anexo 5.4)

La edad al ingreso de los estudiantes oscila entre 17 y 47 años. Se agruparon las edades, según los percentiles, del siguiente modo:

1. De 17 a 21 años
2. De 22 a 23 años
3. De 24 a 25 años
4. Más de 25 años

Las pruebas Log Rank, Breslow y Tarone-Ware rechazan la hipótesis nula de igualdad de funciones de sobrevivencia (□=.05).

Tabla 15. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia

	Estadístico	gl	Significancia
Log Rank	83.93	3	.0000
Breslow	90.30	3	.0000
Tarone-Ware	88.79	3	.0000

Tabla 16. Pruebas de igualdad de funciones de sobrevivencia entre los grupos de edades.

Casos en los que la Ho es rechazada.	Log Rank	Breslow	Tarone – Ware
(2,1)	☑	☑	☑
(3,2)	☑	☑	☑
(4,1)	☑	☑	☑
(4,2)	☑	☑	☑

En la tabla 16 se observa que en el grupo en el que se encuentran los estudiantes cuya edad al ingreso oscila entre 26 y 47 tienen una función de sobrevivencia que presenta diferencias significativas con las funciones de sobrevivencia de los grupo 1 y 2 donde se encuentran los estudiantes más jóvenes Además también hay diferencias significativas entre los dos primeros grupos, es decir, la función de sobrevivencia de los estudiantes que al ingreso tenían entre 17 y 21 años presenta diferencias significativas a la función de sobrevivencia de los estudiantes que al ingreso tenían de 22 a 23 años.

La siguiente tabla 17 muestra la mediana de los tiempos de sobrevivencia de cada uno de estos grupos.

Tabla 17. Medianas del tiempo de sobrevivencia de los diferentes grupos de edad.

Grupos de edad.	Mediana del tiempo de sobrevivencia
1. De 17 a 21 años	4
2. De 22 a 23 años	-
3. De 24 a 25 años	11
4. Más de 25 años	10

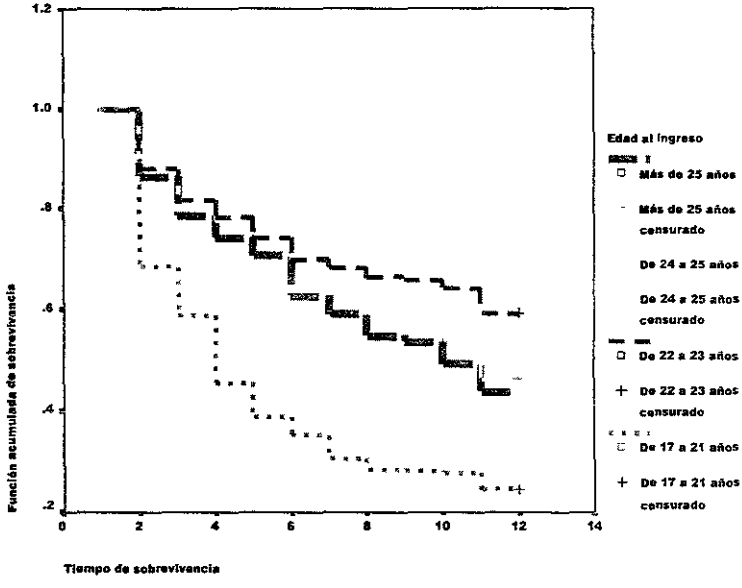
– significa que al doceavo trimestre el evento crítico no había ocurrido en al menos el 50% de las observaciones.

En el grupo en donde se concentran los estudiantes más jóvenes, es decir, el de 17 a 21 años, es el que presenta la menor mediana de tiempo de sobrevivencia El resultado es alarmante ya que al trimestre 4, al menos el 50% de los estudiantes que ingresaron entre la edad de 17 y 21 años fueron no retenidos por la universidad.

En el siguiente grupo formado por los estudiantes que al ingreso tenían entre 22 y 23 años, la mediana del tiempo de sobrevivencia no era aún registrada al cierre de estudio, es decir, al trimestre 12. Resalta el hecho de que en el grupo de más de 36 años, al trimestre 12 tenga una estimación de la mediana del tiempo de sobrevivencia del décimo trimestre y que esta mediana se considerablemente mayor que la del grupo más joven (17 a 21 años).

La gráfica 3 presenta la función de sobrevivencia para cada uno de los grupos de edad.

Gráfica 3. Funciones de sobrevivencia de la variable edad al ingreso.



En la gráfica 3 se observa claramente como el último grupo de edad al ingreso tiene una función de sobrevivencia con valores menores que los de los otros tres grupos. De hecho, del primer al segundo trimestre la función decrece en aproximadamente .03 mientras que para los otros tres grupos de edad el decrecimiento es de aproximadamente .01.

Además, se observa que el grupo en el que se encuentran los estudiantes cuya edad al ingreso oscilaba entre 22 y 23 años, tienen mayor probabilidad de sobrevivir al evento crítico no - retención a partir del séptimo trimestre.

3.4.4.5. Método no paramétrico producto-límite de Kaplan y Meier por escuela de procedencia en el nivel educativo anterior.

(ver anexo 5.5)

La escuela de procedencia en el nivel educativo anterior está dividida en cuatro categorías

1 Colegio de Bachilleres

1. Incorporada a la SEP
2. Otra[1]

Al nivel de significancia de □= 05 no hay diferencias significativas entre las funciones de sobrevivencia de estos 4 grupos

	Estadístico	gl	Significancia
Log Rank	5.14	3	.1616
Breslow	4.32	3	.2293
Tarone-Ware	4.79	3	.1880

El porcentaje de observaciones censuradas está entre el 36 3% y el 48 6% tomando los valores mayores en el grupo de estudiantes cuyo nivel educativo anterior era incorporado a la UNAM o a la SEP, 48.6% y 43 3% respectivamente.

3.4.4.1. Método no paramétrico producto-límite de Kaplan y Meier por promedio en el nivel educativo anterior.

(ver anexo 5.6)

El promedio en el nivel educativo anterior es una variable numérica que oscila entre 6 y 10 Se formaron las siguientes categorías con el fin de comparar las funciones de sobrevivencia de los alumnos con diferentes promedios en la educación media superior.

1. De 6 a 7
2. Más de 7 a 8
3. Más de 8 a 9
4. Más de 9

Al nivel de significancia de □=.05, las pruebas Log Rank, Breslow y Tarone-Ware no rechazan la hipótesis nula de igualdad entre funciones de sobrevivencia de estos 4 grupos.

Tabla 18. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia

	Estadístico	gl	Significancia
Log Rank	7.28	3	.0636
Breslow	6.77	3	.0797
Tarone-Ware	6.97	3	.0729

[1] IPN CECyT, Normal primaria, Universidad Estatal, Incorporada a Universidad Estatal y cualquier otro tipo de Institución que avale certificados de Educación Media Superior

3.4.4.7. Método no paramétrico producto-límite de Kaplan y Meier por examen de admisión.

(ver anexo 5.7)

El examen de admisión de la unidad de CBI de la Universidad Autónoma Metropolitana Unidad Iztapalapa consta de tres partes. La primera se refiere a la habilidad del postulante para resolver problemas de razonamiento matemático, la segunda a la cuantificación del razonamiento verbal del postulante y por último la evaluación de conocimientos generales. El porcentaje de aciertos en el examen de admisión ya sea de manera global, en el área de razonamiento matemático, en el área de razonamiento verbal o en el área de conocimientos se clasifican de en una variable categórica ordinal:

1. Hasta 25%
2. Más de 25% hasta 50%
3. Más de 50% hasta 75%
4. Más de 75%

Las medianas del tiempo de sobrevivencia de los 4 grupos para las variables porcentaje total de aciertos en el examen de admisión, porcentaje de aciertos en razonamiento matemático en el examen de admisión, porcentaje de aciertos en razonamiento verbal en el examen de admisión, porcentaje de aciertos en conocimientos específicos en el examen de admisión, ilustran en la siguiente tabla 19.

Tabla19. Medianas del tiempo de sobrevivencia para porcentaje de aciertos global y en las distintas áreas del examen de admisión.

Medianas del tiempo de sobrevivencia.				
	Hasta 25%	Más de 25% hasta 50%	Más de 50% hasta 75%	Más de 75%
1. Porcentaje total de aciertos en el examen de admisión.	No hay casos	6	11	9
1.1. Porcentaje de aciertos en razonamiento matemático en el examen de admisión.	No hay casos	5	8	11
1.2. Porcentaje de aciertos en razonamiento verbal en el examen de admisión.	11	6	9	-
1.3. Porcentaje de aciertos en conocimientos específicos en el examen de admisión.	4	8	11	11

– significa que al doceavo trimestre el evento crítico no había ocurrido en al menos el 50% de las observaciones

En el caso del porcentaje total de aciertos en el examen de admisión, se observa que en 1995 no ingresaron estudiantes que hayan contestado correctamente menos del 25% del examen de admisión. La mediana más alta del tiempo de sobrevivencia se registra en los estudiantes que tuvieron entre el 50 y 75% de aciertos en el examen de admisión.

A su vez, en 1995 tampoco ingresaron a la universidad alumnos que obtuvieron menos del 25% de aciertos en razonamiento matemático. Las medianas del tiempo de sobrevivencia van aumentando según la categoría. De este modo se observa un valor mayor en la mediana del grupo de estudiantes que obtuvieron más del 75% de aciertos que de los que obtuvieron entre el 25% y 50% o entre el 50% y 75%.

Por otro lado, en el caso del porcentaje de aciertos en razonamiento verbal resalta el hecho de que los alumnos que obtuvieron a lo más el 25% de respuestas correctas, el trimestre 11 sea la mediana de su tiempo de sobrevivencia. Para los alumnos con porcentaje de aciertos entre el 25% y 50% o entre el 50% o 75% el trimestre en el que al menos la mitad de los que ingresaron experimentaron el evento no crítico de no - retención es más temprano, sexto y noveno respectivamente. A su vez, se observa que en los alumnos que contestaron correctamente más del 75%, el evento crítico de no - retención aún no era experimentado por el 50% de los alumnos al cierre del estudio, es decir, al doceavo trimestre.

Por último, en relación con el porcentaje de aciertos en conocimientos generales, se observa que la mediana del tiempo de sobrevivencia es el trimestre 11 para las últimas dos categorías y del cuarto y octavo trimestre para los casos de alumnos con menos de 25% de aciertos y entre 25% y 50% de aciertos respectivamente.

Al nivel de significancia $\alpha = 05$, las pruebas Log Rank, Breslow y Tarone-Ware rechazan la hipótesis nula de igualdad de funciones de sobrevivencia en el caso de porcentaje total de aciertos en el examen de admisión, porcentaje total de aciertos en razonamiento matemático en el examen de admisión y porcentaje total de aciertos en conocimientos generales en el examen de admisión
En el caso de la variable porcentaje de aciertos en razonamiento verbal, la pruebas Tarone-Ware y Log Rank rechazan la hipótesis nula. Sin embrago, la prueba Breslow afirma que no hay diferencias significativas entre las funciones de sobrevivencia

Tabla 20. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia en el caso de porcentaje total de aciertos en el examen de admisión.

	Estadístico	gl	Significancia
Log Rank	29.66	2	.0000
Breslow	33.63	2	.0000
Tarone-Ware	32.19	2	.0000

Tabla 21. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia en el caso de porcentaje total de aciertos en el área de razonamiento verbal en el examen de admisión

	Estadístico	gl	Significancia
Log Rank	11.04	3	.0115
Breslow	7.42	3	.0596
Tarone-Ware	9.26	3	.0261

Tabla 22. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia en el caso de porcentaje total de aciertos en el área de razonamiento matemático en el examen de admisión

	Estadístico	gl	Significancia
Log Rank	10.43	2	.0054
Breslow	10.24	2	.0060
Tarone-Ware	10.52	2	.0052

Tabla 23 Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia en el caso de porcentaje total de aciertos en el área de conocimientos generales en el examen de admisión

	Estadístico	gl	Significancia
Log Rank	23.41	3	.0000
Breslow	30.76	3	.0000
Tarone-Ware	27.45	3	.0000

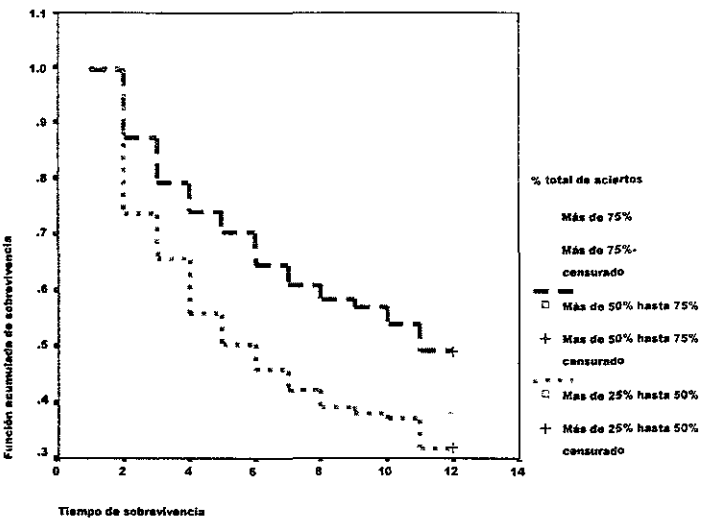
La siguiente tabla 24 presenta las parejas de grupos de porcentaje de aciertos en las tres modalidades del examen de admisión o en puntuación global cuyas hipótesis de igualdad de funciones de sobrevivencia son rechazadas.

Tabla 24. Pruebas de igualdad de funciones de sobrevivencia

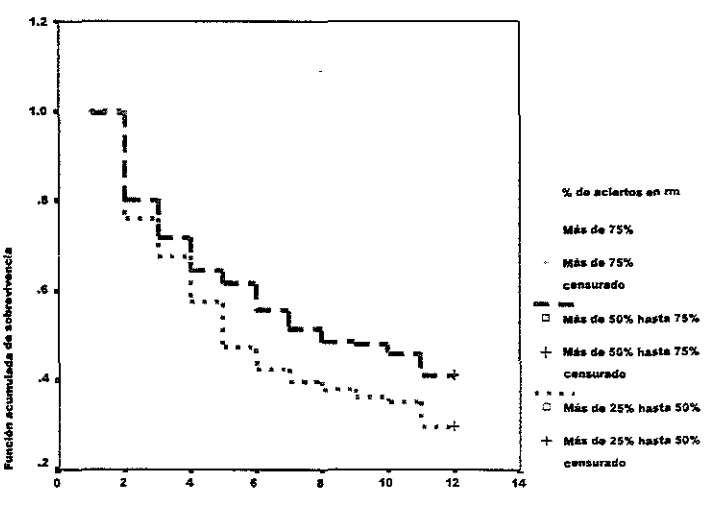
Variable	Casos en los que la Ho es rechazada.	Log Rank	Breslow	Tarone – Ware
Porcentaje total de aciertos total	(3,2)	☑	☑	☑
Porcentaje de aciertos en razonamiento matemático	(3,2)	☑	☑	☑
	(4,2)	☑	☑	☑
Porcentaje de aciertos en razonamiento verbal	(3,2)	☑	---	☑
	(4,2)	☑	---	☑
Porcentaje de aciertos en conocimientos generales.	(2,1)	☑	☑	☑
	(3,1)	☑	☑	☑
	(3,2)	☑	☑	☑
	(4,1)	☑	☑	☑

En la tabla 24 anterior se observa que para las cuatro variables, los grupos que tuvieron entre 25% y hasta 50% de aciertos y más de 50% y hasta 75% tienen funciones de sobrevivencia con diferencias significativas. Además para el caso de razonamiento verbal y razonamiento matemático también hay diferencias significativas entre el grupo de Más de 75% de aciertos y el grupo de Más de 25% y hasta 50% de aciertos. En el caso de conocimientos generales también se observan diferencias significativas entre las funciones de sobrevivencia del primer grupo y los tres restantes. A continuación se presentan las funciones de sobrevivencia de cada una de estas categorías para las cuatro variables relacionadas con el examen de admisión.

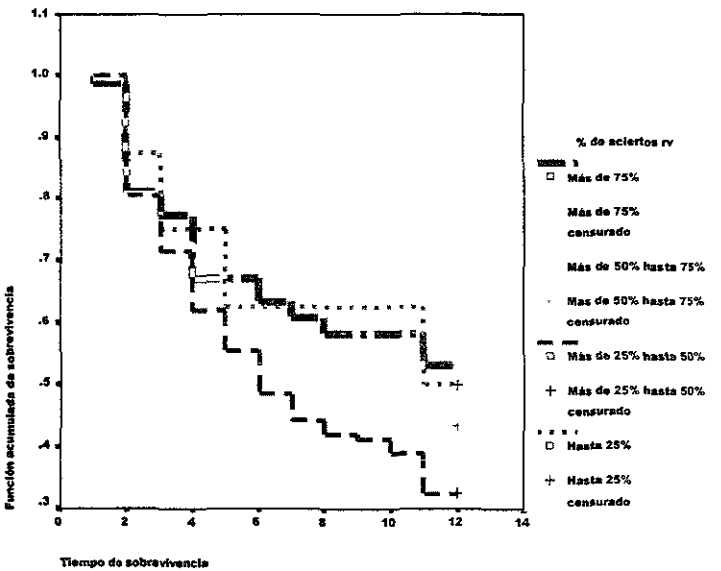
Gráfica 4. Funciones de sobrevivencia de la variable porcentaje total de aciertos en el examen de admisión .



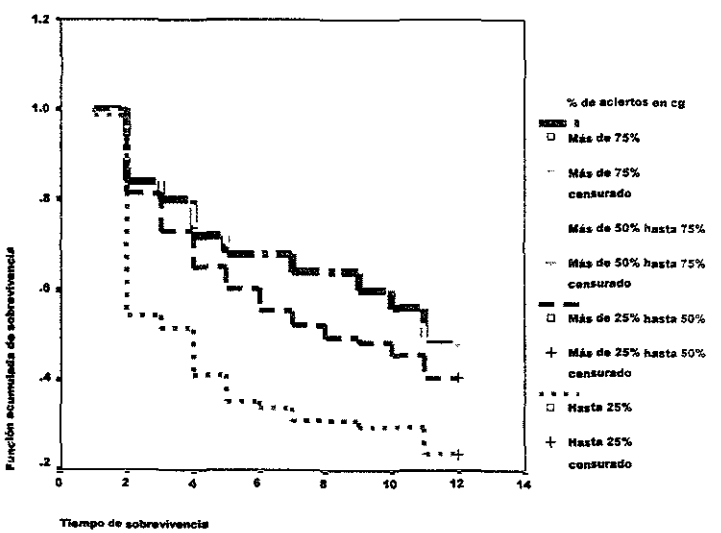
Gráfica 5. Funciones de sobrevivencia de la variable porcentaje de aciertos en el área de razonamiento matemático.



Gráfica 6. Funciones de sobrevivencia de la variable porcentaje de aciertos en el área de razonamiento verbal.



Gráfica 7. Funciones de sobrevivencia de la variable porcentaje de aciertos en el área de conocimientos generales.



3.4.4.8. Método no paramétrico producto-límite de Kaplan y Meier por periodo de ingreso (ver anexo 5 8)

Las medianas del tiempo de sobrevivencia para los alumnos que ingresaron en el periodo de ingreso primavera es el octavo trimestre y el noveno para el caso de los que ingresaron en Otoño.

Al nivel de significancia α=.05, las pruebas Log Rank, Breslow y Tarone-Ware no se rechaza la hipótesis de igualdad de funciones de sobrevivencia.

Tabla 25. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia

	Estadístico	gl	Significancia
Log Rank	.02	1	.8856
Breslow	.74	1	.3888
Tarone-Ware	.25	1	.6141

3.4.4.9. Método no paramétrico producto-límite de Kaplan y Meier por Plan de estudio. (ver anexo 5.9)

La Universidad Autónoma Metropolitana en la división de Ciencias Básica e Ingeniería de la Unidad Iztapalapa imparte nueve diferentes carreras. Al nivel α=.05, las pruebas Log Rank, Breslow y Tarone-Ware rechazan la hipótesis nula de igualdad de funciones de sobrevivencia de las 9 diferentes carreras. La siguiente tabla muestra las parejas de carreras para las cuales la prueba Log Rank rechazo la hipótesis nula de igualdad de funciones de sobrevivencia.

Tabla 26. Pruebas Log Rank, Breslow y Tarone Ware para la comparación de funciones de sobrevivencia

	Estadístico	gl	Significancia
Log Rank	58.90	8	.0000
Breslow	61.51	8	.0000
Tarone-Ware	61.16	8	.0000

Tabla 27. Pruebas de igualdad de funciones de sobrevivencia por carrera.

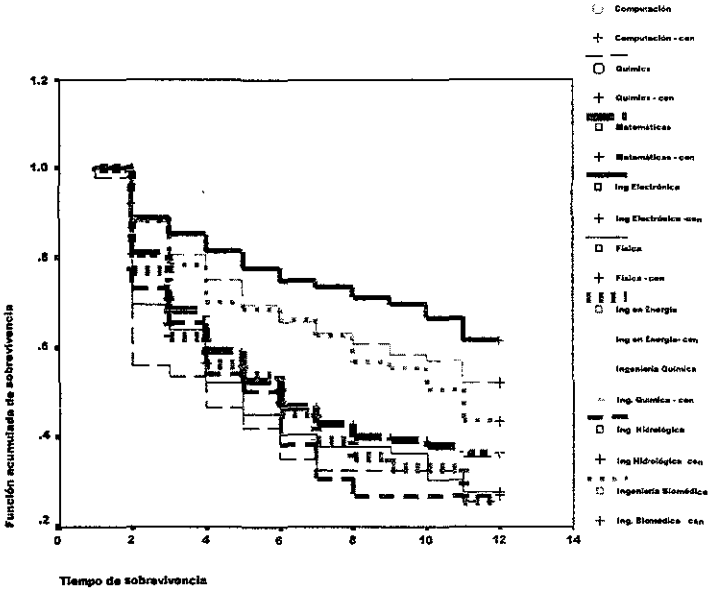
	Ing. Biomédica	Ing. Hidrológica	Ing. Química	Ing. En Energía	Física	Ing. Electrónica	Matemáticas	Química	Ingeniería Computación
Ing. Biomédica	—	☑	----	☑	☑	☑	----	☑	----
Ing. Hidrológica	☑	----	----	----	----	☑	----	----	☑
Ing. Química	----	----	—	----	----	☑	—	☑	☑
Ing. En Energía	☑	—	—	----	----	☑	----	----	☑
Física	☑	—	—	----	----	☑	----	----	☑
Ing. Electrónica	☑	☑	☑	☑	☑		☑	☑	—
Matemáticas	—	—	----	----	—	☑	----	----	—
Química	☑	—	☑	----	----	☑	----	----	—
Ing. Computación	—	☑	☑	☑	☑	----	—	----	—

Se observa en la tabla 27 que la función de sobrevivencia de los estudiantes de Ingeniería Electrónica presenta diferencias significativas con casi todas las demás carreras con excepción de *Ingeniería en Computación. Una posible explicación se puede dar a partir de los resultados* obtenidos en el análisis preliminar del desempeño académico donde se encontró que Ingeniería en Computación e Ingeniería Electrónica eran las dos carreras cuyos alumnos tenían mejor desempeño académico.

La función de sobrevivencia de los alumnos de la Licenciatura en Matemáticas solo presenta diferencias significativas con la función de sobrevivencia de Ingeniería Electrónica

La siguiente gráfica 8 presenta las 9 funciones de sobrevivencia respectivas a cada una de las carreras que imparte la división de Ciencias Básicas e Ingeniería de la Universidad Autónoma Metropolitana , Unidad Iztapalapa.

Gráfica 8. Funciones de sobrevivencia por carrera.



En la gráfica se observa que la función de sobrevivencia de Ingeniería Electrónica toma valores más altos que las funciones de sobrevivencia de las ocho carreras restantes. La segunda y tercera función de sobrevivencia son las correspondientes a Ingeniería Electrónica e Ingeniería Biomédica y, como se muestra en la tabla 27, hay diferencias significativas entre la función de sobrevivencia de Ingeniería Biomédica e Ingeniería Electrónica. Posteriormente se observa la función de sobrevivencia de la Ingeniería en Química la cual reporto diferencias significativas únicamente con Ingeniería Electrónica, Ingeniería en Computación y la Licenciatura en Química.

Las medianas del tiempo de sobrevivencia estiman el trimestre para el cual al menos el 50% de los estudiantes de nuevo ingreso ya había experimentado el evento no-retención. La siguiente tabla 28 presenta dicha información.

Tabla 28. Medianas de los tiempos de sobrevivencia por carrera.

Carrera	Mediana del tiempo de sobrevivencia
1 Ingeniería Biomédica	11
2. Ingeniería Hidrológica	5
3. Ingeniería Química	7
4. Ingeniería en Energía	6
5. Física	5
6. Ingeniería Electrónica	-
7. Matemáticas	6
8. Química	4
9. Ingeniería en Computación	-

– significa que la mediana del tiempo de sobrevivencia todavía no se presentaba al tiempo 12.

En la tabla 28 se observa que por lo menos hasta el trimestre 12, menos del 50% de los estudiantes de Ingeniería Electrónica e Ingeniería en Computación habían sido no - retenidos en la universidad. En el caso de Ingeniería Biomédica, el tiempo en donde se registra al menos el 50% de las no retenciones es el trimestre 11.

Las 7 carreras restantes presentan medianas menores de tiempo de sobrevivencia que oscilan entre el cuarto trimestre en Química hasta el séptimo trimestre en Ingeniería Química

Las carreras de Ingeniería Electrónica e Ingeniería en computación tienen más del 50% de observaciones censuradas. Además, como se menciono anteriormente, en el análisis preliminar de la variable desempeño académico se observó que los alumnos de Ingeniería Electrónica e Ingeniería en Computación tienen mejor desempeño que los alumnos de las seis carreras restantes

3.4.5. Modelo de Cox para datos de sobrevivencia.

El modelo de Cox para datos de sobrevivencia también se le llama *modelo de funciones de riesgo proporcionales de* Cox o *regresión de Cox*. La regresión de Cox establece que la función de riesgo puede ser una función de variables predictoras medidas en escala categórica o de intervalo Por tanto, es más general que el método no paramétrico producto-límite de Kaplan y Meier La regresión de Cox asume que las funciones de riesgo para los diferentes grupos son proporcionales unas a otras a través del tiempo. Si este supuesto no se cumple se utiliza la regresión de Cox con covarianzas variantes

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.

La regresión de Cox es un método semiparamétrico que permite estudiar datos de sobrevivencia, pudiendo resultar de utilidad en muchas situaciones habituales. Modela la función de riesgo de cada individuo como un producto de dos factores:

$$h(t \mid x) = h_0(t)\exp^{\beta_1 x_1 + \ \ + \beta_p x_p}$$

o

$$h(t \mid x) = h_0(t)\exp^{z}$$

El primer factor es una función de riesgo común a todos los individuos, conocida como función de riesgo, y a la que no se le pone ninguna restricción. Esta es la parte no paramétrica del modelo.

El segundo factor es una función de las p covariables del individuo, por lo que tomará un valor distinto para cada individuo. Esta es la parte paramétrica del modelo.

En la regresión de Cox, para la construcción de la función $z = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$, se puede seleccionar el subconjunto de variables independientes que más información aporten sobre la probabilidad de que, para cada posible valor de t, el suceso final no ocurra hasta pasado un periodo de tiempo $t + \Delta t$, supuesto que no ha ocurrido antes de t.

El método Forward, la *Puntuación eficiente de Rao* y el *Estadístico de Wald* son los criterios en los que la Regresión de Cox se basa para la selección y eliminación de variables Dado que se utilizarán en este trabajo, a continuación se describe en general la interpretación de la *Puntuación eficiente de Rao*, del *Estadístico de Wald* y el Método Forward para la selección de variables.

- Estadístico de Wald

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

$$H_0 : \beta_j = 0$$

La interpretación *si dicha hipótesis no se rechaza es que la información que se perdería al eliminar la variable* x_j *no es importante. Si el valor de p asociado al estadístico* Wald *es menor que* α se rechazará la hipótesis nula al nivel de significación α. Bajo este punto de vista , en cada etapa del proceso de selección de variables, la candidata a ser eliminada será la que presente el máximo p-valor asociado al estadístico de Wald. Será eliminada si dicho máximo es mayor que un determinado valor crítico prefijado[1].

- Puntuación eficiente de Rao.

Supongamos que β_j es el parámetro asociado a la variable x_j, supuesto que entrará en la ecuación de regresión en el siguiente paso El estadístico Puntuación eficiente de Rao permite contrastar la hipótesis nula:

$$H_0 : \beta_j = 0$$

La interpretación *si dicha hipótesis no se rechaza es que, si la variable* x_j fuera seleccionada en el siguiente paso, la información que aportaría no sería importante. Si el valor de p asociado al estadístico Puntuación eficiente de Rao es menor que α se rechazará la hipótesis nula al nivel de significación α. Bajo este punto de vista, en cada etapa del proceso de selección

[1] En SPSS es 0.1 pero puede ser modificado.

de variables, la candidata a ser seleccionada será la que presente el mínimo p-valor asociado al estadístico Puntuación eficiente de Rao. Será seleccionada si dicho mínimo es menor que un determinado valor crítico prefijado[2].

- Método hacia delante (forward) para la selección de las variables

Si el proceso comienza sin ninguna variable seleccionada, entonces

1 En el primer paso se introduce la variable que presente el mínimo valor de p asociado al estadístico Puntuación eficiente de Rao, siempre y cuando cumpla el criterio de selección En caso contrario, el proceso finalizará sin que ninguna variable sea seleccionada y, en consecuencia, no será posible construir la función z a partir de la información de las variables independientes.

2. En el segundo paso se introduce la variable que presente el mínimo p-valor asociado al estadístico Puntuación eficiente de Rao. Siempre que se cumpla 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 cumpla el criterio de selección. Si, al introducir una variable, el máximo p-valor asociado al estadístico de Wald para las variables previamente incluidas cumple 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 criterio 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 lo están.

En este trabajo, con el método Forward Wald de la Regresión de Cox, se seleccionarán de las siguientes nueve variables, un subconjunto que aporte más información sobre la probabilidad de que, para cada posible valor de t, el evento crítico no - retención no ocurra hasta pasado un período de tiempo $t + \Delta t$, supuesto que no ha ocurrido antes de t.

[2] En SPSS es 0.05 pero puede ser modificado

1. Género
2. Edad
3. Escuela de procedencia en el nivel educativo anterior
4. Promedio en el nivel educativo anterior
5. Porcentaje de aciertos en el área de matemáticas en el examen de admisión.
6. Porcentaje de aciertos en el área de razonamiento verbal en el examen de admisión.
7. Porcentaje de aciertos en el área de conocimientos generales en el examen de admisión.
8. Periodo de ingreso
9. Carrera (plan de estudio)

Las variables 1,3,8 y 9 son variables categóricas nominales. El paquete estadístico SPSS las introduce al análisis codificándolas automáticamente como variables Dummys.

(Ver anexo 6)

Las siguientes tres tablas ilustran los pasos del Método hacia adelante Wald (Forward Wald) de la Regresión de Cox.

Tabla 29. Variables que no están en la ecuación

	Puntaje	gl	Sig.
INGRESO	.018	1	.892
PLA	52.106	8	.000
PLA(1)	.654	1	.419
PLA(2)	2.921	1	.087
PLA(3)	.384	1	.535
PLA(4)	8 455	1	004
PLA(5)	7.302	1	.007
PLA(6)	23.320	1	.000
PLA(7)	2.421	1	.120
PLA(8)	7.808	1	.005
EDAD	23.191	1	000
SEX	2.708	1	.100
PRV	9.535	1	.002
PRM	13.980	1	.000
PCO	22.278	1	.000
ESCUELA	4.556	3	.207
ESCUELA(1)	1.497	1	.221
ESCUELA(2)	2.348	1	.125
ESCUELA(3)	.797	1	.372
PROMEDIO	4.013	1	045

a Residual Ji-Cuadrado = 104.144 con 18 gl Sig. = .000

b Número del bloque de inicio 0, función inicial Log verosimilitud : -2 Log verosimilitud· -5865 176

Tabla 30. Variables en la ecuación

		B	SE	Wald	df	Sig.	Exp(B)
Paso 1	PLA			49 724	8	000	
	PLA(1)	189	.195	942	1	332	1.208
	PLA(2)	.692	264	6 895	1	009	1.998
	PLA(3)	.381	187	4 160	1	041	1 464
	PLA(4)	.658	182	13 072	1	000	1.931
	PLA(5)	.672	.192	12.267	1	.000	1 957
	PLA(6)	- 291	189	2 360	1	.124	747
	PLA(7)	458	168	7 399	1	007	1 581
	PLA(8)	785	.219	12 836	1	000	2.193
Paso 2	PLA			51 209	8	000	
	PLA(1)	.146	195	560	1	.454	1 157
	PLA(2)	.724	264	7 535	1	006	2 062
	PLA(3)	.322	.187	2.962	1	.085	1 380
	PLA(4)	.678	.182	13 868	1	000	1 971
	PLA(5)	708	.192	13.606	1	000	2 030
	PLA(6)	- 306	.190	2 607	1	106	.736
	PLA(7)	.433	169	6 612	1	010	1 543
	PLA(8)	.702	.220	10 206	1	001	2 018
	EDAD	-.070	014	24 045	1	.000	.932
Paso 3	PLA			53.987	8	000	
	PLA(1)	205	196	1 095	1	295	1.227
	PLA(2)	687	264	6 778	1	.009	1.987
	PLA(3)	.366	187	3 806	1	.051	1 441
	PLA(4)	.647	182	12 581	1	000	1 909
	PLA(5)	.718	.192	13.988	1	000	2 050
	PLA(6)	- 346	.190	3 323	1	068	708
	PLA(7)	.455	169	7.262	1	007	1 576
	PLA(8)	.792	221	12.812	1	000	2 208
	EDAD	- 078	015	28 575	1	000	925
	SEX	.418	128	10 587	1	001	1 518

Tabla 36. Variables que no están en la ecuación

		Puntaje	gl	Sig.
Paso 1	INGRESO	113	1	737
	EDAD	24 120	1	000
	SEX	5 508	1	019
	PRV	1 187	1	276
	PRM	232	1	630
	PCO	3 640	1	056
	ESCUELA	4 615	3	202
	ESCUELA(1)	973	1	324
	ESCUELA(2)	2 168	1	141
	ESCUELA(3)	.833	1	361
	PROMEDIO	2 084	1	149
Paso 2	INGRESO	3 111	1	078
	SEX	10 720	1	001
	PRV	1 322	1	.250
	PRM	.290	1	590
	PCO	066	1	.797
	ESCUELA	7 320	3	062
	ESCUELA(1)	2.823	1	093
	ESCUELA(2)	4 171	1	041
	ESCUELA(3)	940	1	332
	PROMEDIO	5 695	1	017
Paso 3	INGRESO	2 516	1	113
	PRV	1.129	1	288
	PRM	112	1	737
	PCO	189	1	664
	ESCUELA	6 654	3	084
	ESCUELA(1)	2 587	1	108
	ESCUELA(1)	2.587	1	.108
	ESCUELA(2)	3.131	1	077
	ESCUELA(2)	3.131	1	077
	ESCUELA(3)	1 373	1	.241
	ESCUELA(3)	1.373	1	241
	PROMEDIO	3.159	1	075
	PROMEDIO	3.159	1	075

a Residual Ji-cuadrado = 49.201 con 10 df Sig = 000
b Residual Ji-cuadrado = 25.876 con 9 df Sig. = 002
c Residual Ji-cuadrado = 15 160 con 8 df Sig. = .056

En la tabla 29 se observa que la estadístico Puntuación eficiente de Rao más alto y por tanto el mínimo valor de p asociado es el de la variable PLA= Carrera , entonces es la primera variable que entra al subconjunto de variables que mejor explican la probabilidad de que el evento crítico de no - retención aparezca.

En la tabla 30 se verifica por medio del estadístico *Wald* que la variable PLA=Carrera debe ser incluida en el modelo. Posteriormente en la tabla 31 se observa (en el paso 1) que la siguiente variable cuya valor estimado del estadístico *Puntuación eficiente de Rao* es el más alto, es el correspondiente a la variable edad y en la tabla 30 que el estadístico *Wald* rechaza la hipotesis nula, por tanto, en el paso 2, la variable edad entra en el modelo. En el siguiente paso, la variable

cuya estadístico *Puntuación eficiente de Rao* (tabla 31, paso 2) es más alto es la variable género y también el estadístico *Wald* (tabla 30, paso 3) rechaza la hipótesis nula por tanto la variable género también entra en el modelo En el cuarto paso, la variable que entraría sería Escuela sin embargo el estadístico *Puntuación eficiente de Rao* no rechaza la hipótesis nula $H_0 : \beta_j = 0$ Por tanto las variables que mejor explican la probabilidad de que el evento crítico no-retención aparezca son el plan de estudios, la edad al ingreso y el género.

Tabla 32. Prueba Omnibus de Coeficientes del Modelo

	-2 Log de la verosimilitud	Overall (puntaje)			Cambio del anterior paso			Cambio del anterior bloque		
Paso		Ji-cuadrada	Gl	Sig.	Ji-cuadrada	gl	Sig.	Ji-cuadrado	gl	Sig.
1	5812 292	52 106	8	000	52 884	8	000	52 884	8	.000
2	5784 362	78.077	9	.000	27.930	1	000	80.813	9	000
3	5772.952	89.681	10	.000	11.410	1	.001	92 224	10	000

a Variable(s) ingresadas al Paso número 1 PLA
b Variables ingresadas al Paso número 2: EDAD
c Variables ingresadas al Paso número 3. SEX
d Primer Bloque número 0, función inicial Log de la verosimilitud -2 Log de la verosimilitud 5865 176
e Primer Bloque número 1. Metodo Forward Stepwise (Wald)

En la tabla 32 anterior se observa, como ya se mencionó, que las variables independientes seleccionados que aportan más información sobre la probabilidad de que el evento crítico no - retención ocurra son : *carrera, edad al ingreso y género.*

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

$$-2LL = -2Log_e \text{ de la verosimilitud}$$

Cuanto más próximo a 0 sea el valor del estadístico $-2LL$[1], más próxima a 1 será la verosimilitud y mejor será el modelo.

En la tabla 32 se observa que el estadístico $-2LL$ decreció en cada uno de los tres pasos. El valor inicial de $-2LL$ es de –5865.176. En el primer paso este valor disminuye a 5812 292 produciéndose un cambio de 52.884 respecto al paso anterior. Es decir, antes de introducir ninguna variable independiente el valor del estadístico $-2LL$ era mayor. La razón de la

existencia de este valor radica en que el modelo que se esta validando es el correspondiente a la función $h(t \mid x)$. Cuando no se consideran los valores de las variable independientes o, lo que es equivalente , cuando son todos iguales a cero, al ser la función z igual a 0, $\exp^z$ = 1 y, en consecuencia , $h(t \mid x)$ se estima considerando únicamente el tiempo de sobrevivencia observado en cada caso, independientemente de los valores de las variables independientes. Luego el cambio detectado en el estadístico $-2LL$ en este primer paso permitirá evaluar la mejora que se produce en el modelo al incorporar la variable carrera. El p-valor asociado (Sig = 0 0000) es menor que 0.05, luego al nivel de significación 0 05, se puede aceptar que el cambio es estadísticamente significativo.

En el segundo y tercer paso se observa también que $-2LL$ disminuye y que el valor de p es menor 0.05, entonces se puede aceptar que los cambios también son estadísticamente significativos.

Entonces regresando a nuestra ecuación de regresión original tenemos,

$$h(t \mid x) = h_0(t)\exp^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}$$

que es equivalente a ,

$$h(t \mid x) = h_0(t)\exp^{\beta_1 x_1} \exp^{\beta_2 x_2} \exp^{\beta_3 x_3}$$

donde x_1 corresponde a la carrera, x_2 a la variable edad al ingreso y x_3 al género del estudiante. En la tabla 30, en la columna $\exp(\beta)$ se presenta los valores correspondientes a $\exp^{\beta_i}$

Ahora supongamos que queremos calcular $h(t \mid x)$ para los estudiantes, hombres y mujeres, de la Licenciatura en Matemáticas y cuya edad al ingreso fue de 19 años.

Entonces, para el caso de estudiantes mujeres

$$h(t \mid x) = h_0(t)1.576^1 * 0.925^{19} * 1.518^0 = h_0(t)0.3583$$

[1] LL son las siglas en inglés LOGLIKELIHOOD

y para el caso de estudiantes hombres,

$$h(t \mid x) = h_0(t)1.576^1 * 0.925^{19} * 1.518^1 = h_0(t)0.5439$$

En otras palabras, si el estudiante ingresa a la Licenciatura en Matemáticas y su edad al ingreso era de 19 años, para cualquier t, la probabilidad estimada de que el alumno sea no retenido por la universidad en un pequeño intervalo $(t, t+\Delta t)$ (en el límite cuando Δt tiende a 0), supuesto que no ha reaparecido antes del instante t, es mayor en el caso de los hombres que en el caso de las mujeres.

En las siguientes tres tablas 33,34 y 35 se presentará la función $h(t \mid x)$ para las 9 carreras, ambos géneros y suponiendo edad al ingreso de 21, 22 y 23 años respectivamente Se eligieron estas edades de ingreso porque son en las edades más frecuentes en los alumnos al ingreso

Tabla 33. Probabilidad de experimentar el evento no - retención para hombres y mujeres de las diferentes carreras y cuya edad al ingreso fue de 21 años.

Carrera	Género	
	$h(t \mid x)$ Hombres	$h(t \mid x)$ Mujeres
1. Ingeniería Biomédica	$h_0(t)$ 0.362	$h_0(t)$ 0 239
2 Ingeniería Hidrológica	$h_0(t)$ 0.587	$h_0(t)$ 0 387
3. Ingeniería Química	$h_0(t)$ 0 426	$h_0(t)$ 0 280
4 Ingeniería en Energía	$h_0(t)$ 0.564	$h_0(t)$ 0.371
5. Física	$h_0(t)$ 0 605	$h_0(t)$ 0 399
6. Ingeniería Electrónica	$h_0(t)$ 0 209	$h_0(t)$ 0.138
7. Matemáticas	$h_0(t)$ 0 465	$h_0(t)$ 0.307
8 Química	$h_0(t)$ 0 652	$h_0(t)$ 0 430
9 Ingeniería en Computación	$h_0(t)$ 0 295	$h_0(t)$ 0 195

En los nueve planes de estudio, se observa que para los estudiantes cuya edad de ingreso fue de 21 años, los hombres tienen mayor probabilidad de ser no-retenidos por la universidad que las mujeres. Además las carreras con más riesgo son Física, Química e Ingeniería Hidrológica

mientras que las de menor riesgo son Ingeniería Electrónica , Ingeniería en Computación e Ingeniería Biomédica.

Tabla 34. Probabilidad de experimentar el evento no - retención para hombres y mujeres de las diferentes carreras y cuya edad al ingreso fue de 22 años.

Carrera	Género	
	$h(t \mid x)$ Hombres	$h(t \mid x)$ Mujeres
1 Ingeniería Biomédica	$h_0(t)$ 0 335	$h_0(t)$ 0 221
2. Ingeniería Hidrológica	$h_0(t)$ 0 543	$h_0(t)$ 0 358
3 Ingeniería Química	$h_0(t)$ 0 394	$h_0(t)$ 0 259
4. Ingeniería en Energía	$h_0(t)$ 0 521	$h_0(t)$ 0 343
5. Física	$h_0(t)$ 0 560	$h_0(t)$ 0.369
6. Ingeniería Electrónica	$h_0(t)$ 0 193	$h_0(t)$ 0 127
7. Matemáticas	$h_0(t)$ 0.430	$h_0(t)$ 0 284
8 Química	$h_0(t)$ 0.603	$h_0(t)$ 0 397
9 Ingeniería en Computación	$h_0(t)$ 0 273	$h_0(t)$ 0 180

Tabla 35. Probabilidad de experimentar el evento no - retención para hombres y mujeres de las diferentes carreras y cuya edad al ingreso fue de 23 años.

Carrera	Género	
	$h(t \mid x)$ Hombres	$h(t \mid x)$ Mujeres
1. Ingeniería Biomédica	$h_0(t)$ 0.310	$h_0(t)$ 0 204
2. Ingeniería Hidrológica	$h_0(t)$ 0.502	$h_0(t)$ 0 331
3. Ingeniería Química	$h_0(t)$ 0.364	$h_0(t)$ 0 240
4. Ingeniería en Energía	$h_0(t)$ 0 482	$h_0(t)$ 0.318
5 Física	$h_0(t)$ 0 518	$h_0(t)$ 0 341
6 Ingeniería Electrónica	$h_0(t)$ 0 179	$h_0(t)$ 0 118
7. Matemáticas	$h_0(t)$ 0 398	$h_0(t)$ 0.262
8. Química	$h_0(t)$ 0.558	$h_0(t)$ 0 368
9. Ingeniería en Computación	$h_0(t)$ 0 253	$h_0(t)$ 0 166

En las tablas 33 y 34 se observa que también en los estudiantes cuya edad al ingreso fue de 22 o 23 años, la probabilidad de experimentar el evento crítico no - retención es mayor en hombres que en mujeres. Además las carreras con menor riesgo son, como en el caso de estudiantes cuya edad al ingreso fue de 21 años, Ingeniería Electrónica, Ingeniería en Computación e Ingeniería Biomédica.

También se observa que los estudiantes cuya edad al ingreso fue de 23 años tienen menos probabilidad de experimentar el evento crítico no - retención que los estudiantes que ingresaron de 21 o 22 años. Si recordamos, en los resultados obtenidos con el método producto-límite Kaplan y Meier teníamos que los estudiantes cuya edad al ingreso oscilaba entre 17 y 21 años eran quienes tenían probabilidad más alta de ser no retenidos en la universidad en comparación con los estudiantes cuya edad al ingreso era mayor. De hecho se observaba, que a partir del séptimo trimestre los estudiantes cuya edad al ingreso era de 22 a 23 años eran quienes tenían mayor probabilidad de sobrevivencia al estado crítico no - retención.

En general, de esta manera puede calcularse la probabilidad, en función del tiempo, de que un estudiante sea no retenido por la universidad dado su género, edad al ingreso y carrera que cursa.

CAPÍTULO 4
DISCUSIÓN

El propósito original de este trabajo era el estudiar la deserción estudiantil, sin embargo como se expuso en el apartado 2.2 del capítulo 2, la información de la que se dispone no permite abordar este problema como tal. El principal problema radica en que el Reglamento General de Estudios de la Universidad Autónoma Metropolitana establece como plazo máximo de permanencia en la universidad 10 años que equivalen a 30 trimestres. Además, la legislación establece que un alumno puede abandonar sus estudios a lo más por un periodo de 6 trimestres Si este periodo de abandono se cumple, entonces el alumno debe presentar un examen de conocimientos de las UEA's acreditadas antes de este abandono, durante los seis periodos lectivos desde el abandono. Cabe mencionar que este recurso, en general, no es utilizado. Entonces, para determinar si un alumno es desertor sería necesario contar con información de diez años desde su ingreso porque con la información disponible, doce trimestres, no es posible determinar si los alumnos que abandonaron sus estudios desde el séptimo trimestre regresarán en algún momento a la universidad. Por tanto, para poder abordar el fenómeno de la deserción estudiantil en la universidad Autónoma Metropolitana y poder considerar a un alumno desertor, sería necesario contar con información de 10 años de su trayectoria escolar desde su ingreso

Por las razones anteriores, surge la necesidad de un término alterno que cubra las expectativas del término deserción estudiantil Es así como nace el término - retención *estudiantil. Por definición, un alumno es retenido en la universidad si habiéndose inscrito en un periodo lectivo* dado se inscribe en el consecutivo. Sin embargo, con este nuevo concepto, nos enfrentamos a otras limitaciones dadas las características de la universidad Pensemos en los casos de alumnos que abandonan sus estudios por un trimestre o dos, por lo que deciden no inscribirse en dichos trimestres, pero en los trimestres posteriores a éstos regresan a la universidad. Si nos apegáramos estrictamente a la definición de alumno retenido, estos alumnos serían no retenidos en la universidad lo cual no reflejaría la situación real de dichos alumnos. Debido a ello, en este estudio y con base en la legislación de la universidad, un alumno es no retenido en la universidad si abandona sus estudios por más de seis trimestres consecutivos. Por otro lado, dado que un alumno tiene hasta 30 trimestres para cubrir el 100% de créditos, entonces un alumno será también considerado no retenido en la universidad si no abandona sus estudios por más de seis periodos consecutivos, pero si su desempeño académico acumulado al doceavo trimestre, es decir, su avance de créditos acumulados, no le permitiera cubrir, en los 18 trimestres consecutivos el 100% de créditos al ritmo de acreditación que, según las evidencias, tiene la capacidad de lograr Por ejemplo, supongamos un estudiante de la Licenciatura en Física, que consta de 480 créditos, entonces si el alumno no abandonó sus estudios por más de seis trimestres consecutivos pero su

desempeño académico acumulado al doceavo trimestre es menor que (550 / 30) * 12 , entonces es un alumno no retenido en la universidad

El término *retención estudiantil* tiene una relación inversa con el término *deserción estudiantil*. A mayor deserción en la universidad, menor retención en la misma. Del mismo modo, a menor deserción mayor retención

- Etapa 1: Análisis preliminar del desempeño académico.

Hasta ahora el término desempeño académico sólo ha sido utilizado para determinar si un alumno es retenido o no en la universidad. El término desempeño académico también permite abordar la retención estudiantil bajo el supuesto de que un alumno con buen desempeño académico tiene mayor probabilidad de permanecer en la universidad y, por tanto, menor probabilidad de desertar. Es entonces como surge la idea de analizar el desempeño académico de los estudiantes por carrera y periodo de ingreso (Apartado 3.2 del capítulo 3). En principio, se observó que el desempeño académico de los estudiantes de la misma carrera era muy dispar y también entre las carreras. Sin embargo, en general se observó que el desempeño académico de los estudiantes de las carreras de Ingeniería Electrónica e Ingeniería en Computación es mejor que en las 7 carreras restantes. Además, se observó que el desempeño académico de los estudiantes que ingresaban en el periodo otoño era mejor que el de aquellos que ingresaban en el periodo primavera. Esto sugiere que, posiblemente, las licenciaturas de Ingeniería en Computación y en Ingeniería Electrónica tienen mayor inserción en el campo de trabajo, por lo que el estudiante está más motivado a finalizar sus estudios universitarios y que, posiblemente, los estudiantes que ingresan en otoño tienen mejor desempeño académico porque, en general, son alumnos con trayectoria escolar continua, ya que ingresaron a la universidad al finalizar su educación media superior y no tuvieron que esperar, al menos un semestre, como los estudiantes que ingresaron en el periodo primavera.

La diferencia observada en el desempeño académico de los estudiantes dentro de las mismas carreras y entre ellas, sugirió la siguiente etapa de la estrategia de análisis. Resultaba de interés encontrar las características que determinaban que un estudiante tuviera mejor desempeño que otro, para lo cual se utilizó el método no jerárquico de k medias del análisis de conglomerados. Este método fue elegido por el tamaño de población con el que se cuenta y para el cual los métodos jerárquicos resultan confusos en cuanto a su interpretación. Se determinó que se clasificaría a los estudiantes en 5 grupos, de acuerdo con sus características, tales como edad al ingreso, promedio en el nivel educativo anterior y porcentaje de aciertos en razonamiento matemático, conocimientos generales y razonamiento verbal del examen de admisión. Las

variables género y escuela de procedencia en el nivel educativo anterior no fueron incluidas para el análisis de conglomerados, dado que el método no jerárquico de k-medias utiliza la distancia euclidiana como medida de similaridad. Sin embargo, después de formados los conglomerados se incluyeron en el análisis.

- Etapa 2: Análisis de conglomerados

El objetivo de esta etapa fue modelar el desempeño académico de los cinco grupos formados mediante el análisis de conglomerados, con el objeto de identificar y encontrar las características que determinan que un estudiante tenga mejor desempeño académico que otro. Mediante los conglomerados se etiquetaron jerárquicamente del siguiente modo:

1. Mayor potencial de éxito
2. Aceptable potencial de éxito
3. Moderado potencial de éxito
4. Bajo potencial de éxito
5. Casi nulo potencial de éxito

En general, se observó que, a pesar de que los estudiantes más jóvenes se encuentran clasificados en el "conglomerado de mayor potencial de éxito", la diferencia de edades de estos estudiantes y de los que integran los conglomerados "bajo potencial de éxito" y "casi nulo potencial de éxito" es muy pequeña. En el conglomerado de "moderado potencial de éxito" se encuentran los estudiantes más grandes ya que al ingreso tenían, en su mayoría, más de 27 años de edad. En el conglomerado "aceptable potencial de éxito", la mayoría de los estudiantes tenían una edad al ingreso entre 22 y 26 años.

En relación con el examen de admisión, se observó que en los conglomerados "mayor potencial de éxito", "aceptable potencial de éxito" y "moderado potencial de éxito", el porcentaje de estudiantes que obtuvieron más del 75% de aciertos en el área de razonamiento matemático es casi la mitad o más de la mitad. Para los dos últimos conglomerados, el porcentaje de estudiantes que obtuvieron más del 75% es casi nulo. Los resultados en esta área del examen de admisión son mejores que en el área de conocimientos generales y que en el área de razonamiento verbal. No hay estudiantes en ninguno de los cinco conglomerados cuyo porcentaje de aciertos en esta área sea menor del 25%.

En lo que respecta al área de conocimientos generales del examen de admisión, se observó que los estudiantes del conglomerado "moderado potencial de éxito" son quienes tuvieron un mejor desempeño. Una posible explicación es que en este conglomerado se encuentran los estudiantes cuya edad al ingreso, en general, era de 27 años en adelante, por lo que la adquisición de conocimientos generales responde al proceso natural de la experiencia. Los resultados en esta área también son mejores para los dos primeros conglomerados que para los dos últimos, como sucede en el caso del área de razonamiento matemático.

En relación con los resultados obtenidos en el área de razonamiento verbal del examen de admisión, no se observa una tendencia específica en los conglomerados. El desempeño en esta área del examen es un tanto similar para los conglomerados "mayor potencial de éxito", "aceptable potencial de éxito", "moderado potencial de éxito" y "casi nulo potencial de éxito". En el conglomerado "bajo potencial de éxito" se observó que el porcentaje promedio de aciertos es sensiblemente menor que el de los otros cuatro conglomerados.

En el promedio del nivel educativo anterior, los integrantes del conglomerado "mayor potencial de éxito" alcanzaron promedios más altos que los estudiantes de los otros cinco conglomerados. El conglomerado cuyos estudiantes tienen promedio más bajo es el de "aceptable potencial de éxito". La media del promedio en el nivel educativo anterior del conglomerado "moderado potencial de éxito" es la segunda más alta después de la del conglomerado "mayor potencial de éxito". La mayoría de los estudiantes que integran los conglomerados "bajo potencial de éxito" y "casi nulo potencial de éxito" finalizaron su educación media superior con un promedio inferior a ocho.

Hasta aquí se logró agrupar a los estudiantes en 5 conglomerados con características homogéneas y modelar el desempeño académico de los estudiantes que componen cada uno de estos cinco conglomerados.

Como se mencionó en el apartado 2.3 del capítulo 2, éste es un estudio longitudinal ya que se cuenta con la información del avance de créditos por trimestre, durante 12 periodos, de los estudiantes que integran la población objetivo. La siguiente etapa de la estrategia del análisis consiste en introducir la técnica de análisis de sobrevivencia, que resulta muy útil en estudios longitudinales ya que permite estimar la probabilidad, en función del tiempo, de que el evento crítico de interés ocurra. Como ya se mencionó, el evento crítico de interés es la no - retención, es decir, se estima la probabilidad, en función del tiempo, de que un alumno sobreviva al evento crítico no - retención.

En este estudio, la técnica de análisis de conglomerados fue útil siendo utilizada complementariamente con el análisis de sobrevivencia ya que si sólo nos hubiéramos limitado a llevar a cabo el análisis de conglomerados, el alcance que hubiésemos obtenido sería tan solo homogeneizar 5 grupos y después etiquetarlos modelando el desempeño académico de los estudiantes que los componen. Además, como se mencionó, el análisis de conglomerados tiene sus limitaciones en el sentido de que, al ser una población de tamaño grande, se utilizó el método no jerárquico de k medias, el cual tiene la desventaja latente de no integrar variables categóricas La ventaja de utilizar ambas técnicas complementariamente es que es posible estimar la probabilidad de que un estudiante perteneciente a un conglomerado dado, sea no retenido en la universidad, en función del tiempo.

- Etapa 3. Análisis de sobrevivencia

La etapa del análisis de sobrevivencia se divide en dos sub-etapas. La primera se refiere al método no paramétrico producto-límite de Kaplan y Meier, que permite estimar la función de sobrevivencia de uno o más grupos y después, por medio de las pruebas Log Rank, Breslow o Tarone Ware, establecer si hay diferencias significativas entre las funciones de sobrevivencia de estos grupos. La segunda sub-etapa consiste en el modelo de regresión de Cox, el cual permite calcular la probabilidad de ocurrencia del evento crítico, en función del tiempo, integrando todas las variables de interés y seleccionando las más importantes.

- Sub-etapa 1 del análisis de sobrevivencia: Método no paramétrico producto-límite de Kaplan y Meier.

En la primera sub-etapa se estimó la función de sobrevivencia por conglomerado, por conglomerado estratificado por género y escuela de procedencia, por género, por edad al ingreso, por promedio en el nivel educativo anterior, por porcentaje total de aciertos en el examen de admisión, por porcentaje de aciertos en el examen de admisión en el área de razonamiento matemático, por porcentaje de aciertos en el examen de admisión en el área de razonamiento verbal, por porcentaje de aciertos en el examen de admisión en el área de conocimientos generales, por periodo de ingreso y por carrera.

En principio, los resultados que se obtuvieron en esta sub-etapa permitieron verificar el supuesto de que a mejor desempeño académico, mayor probabilidad de ser retenido en la universidad.

Estimar las funciones de sobrevivencia permitió verificar las etiquetas asignadas a los conglomerados, las cuales resultaron acertadas salvo en el caso de los dos últimos conglomerados, ya que a partir del cuarto trimestre los integrantes del conglomerado casi nulo potencial de éxito, tienen mayor probabilidad de sobrevivir que los integrantes del conglomerado bajo potencial de éxito.

En la estimación de las funciones de sobrevivencia para conglomerados estratificados por género y escuela de procedencia se observó que sólo para los estudiantes de género masculino, la circunstancia de haber terminado la educación media superior en escuelas incorporadas a la SEP influye en su probabilidad de sobrevivencia En el caso de las mujeres no influye. Además, se observó que, a excepción de las estudiantes que obtuvieron su certificado de educación media superior de escuelas clasificadas como "Otras", las mujeres tienen mayor probabilidad de permanecer en la Institución que los hombres A su vez, el porcentaje de casos censurados para hombres y mujeres, es mayor en el caso de los estudiantes que terminaron su educación media superior en escuelas incorporadas a la UNAM, que en cualquiera de las otras tres clasificaciones

Al realizar el análisis de sobrevivencia por medio del método producto-límite de Kaplan y Meier por variable, se obtuvieron algunas diferencias en los resultados. Se observó que en el caso de la variable género, las funciones de sobrevivencia no tienen diferencias significativas. A su vez, en el caso de la variable edad al ingreso, se observó que los estudiantes más jóvenes (de 17 a 21) son quienes tienen menor probabilidad de sobrevivencia En este caso, las pruebas Log Rank, Breslow y Tarone Ware determinaron diferencias significativas entre los grupos de dicha variable, estableciendo que la función de sobrevivencia del grupo más joven toma valores más pequeños que las de los otros grupos.

En relación con la escuela de procedencia se observa que los alumnos egresados de escuelas incorporadas a la UNAM tienen mayor probabilidad de permanecer en la Institución que aquellos que obtuvieron su certificado de Educación Media Superior en Instituciones referentes a las tres categorías restantes. Por ejemplo, este resultado es igual al obtenido en los conglomerados estratificados por género y escuela de procedencia, donde se observó que el porcentaje de hombres y mujeres que terminaron su educación media superior en escuelas incorporada a la UNAM era mayor que el porcentaje de observaciones censuradas en los otros tres tipo de Instituciones Por su lado, el segundo tipo de Institución de egreso de Educación Media Superior que presenta probabilidad más alta de sobrevivencia son las escuelas incorporadas a la SEP. El promedio de egreso de la EMS tampoco presenta diferencias significativas entre las funciones de sobrevivencia.

El área del examen de admisión que más determina alta probabilidad de retención es el área de razonamiento matemático seguida del área de conocimientos generales y, finalmente, el área de razonamiento verbal Las carreras en las que se observa mayor probabilidad de retención son Ingeniería Electrónica e Ingeniería en Computación y en las que menos se observa son Ingeniería en Energía, Ingeniería Hidrológica, Química y Física A su vez, las funciones de sobrevivencia del periodo de ingreso a la universidad no presentan diferencias significativas, es decir, el periodo de ingreso parece no determinar la permanencia o no de los estudiantes en la universidad como se había supuesto en el análisis preliminar del desempeño académico por carrera.

Por tanto, los estudiantes con mayor probabilidad de sobrevivir al evento crítico no - retención son las mujeres que terminaron su educación media superior en escuelas incorporadas a la UNAM, con promedio mayor de siete y cuya edad al ingreso fluctuaba alrededor de 22 años Es deseable que en el examen de admisión, en el área de razonamiento matemático obtenga más del 75% de aciertos, en el área de conocimientos generales más del 50% y en razonamiento verbal de preferencia un porcentaje alto de aciertos.

- Sub-etapa 2 del análisis de sobrevivencia Modelo de Cox para datos de sobrevivencia

Como ya se mencionó, el análisis de sobrevivencia comprende dos sub-etapas La segunda sub-etapa consiste en el modelo de regresión de Cox, con el cual se calcula la probabilidad de ocurrencia del evento crítico, en función del tiempo, integrando todas las variables de interés y seleccionando las más importantes. En esta sub-etapa, con el método paso a paso Wald (forward Wald) se seleccionaron las variables que más influyen en la probabilidad de que el evento crítico no - retención ocurra. Se encontró que dichas variables son la carrera, la edad al ingreso, y el género.

Como se observó con Kaplan Meier, las carreras en las que se observó menor ocurrencia del estado crítico no - retención fueron Ingeniería en Computación e Ingeniería Electrónica Así mismo, también se determinó que las mujeres son las que tienen mayor probabilidad de éxito que los hombres y que, en relación con la edad de ingreso, sorprendentemente los alumnos más jóvenes no son los que tienen más probabilidad de sobrevivir al evento crítico no - retención, sino los que ingresan de 22 años en adelante. Estos resultados obtenidos con el método no paramétrico producto-límite de Kaplan y Meier se confirman, ya que se observa que en general las mujeres, de cualquier edad al ingreso y de cualquier carrera, tienen mayor probabilidad de sobrevivencia que los hombres. Así mismo se observó que las carreras cuyos alumnos tienen mayor probabilidad de permanecer en la Institución son Ingeniería en Computación e Ingeniería Electrónica

- Recomendaciones y limitaciones

Definitivamente, el alcance de este estudio es limitado dado que sólo se cuenta con información relativa a dos dimensiones, características individuales y antecedentes educativos, de las siete que Vincent Tinto plantea. Sin embargo, al principio se pensó en integrar variables socioeconómicas, lo cual no fue posible por la falta de información organizada por parte de la universidad y que hubiese permitido conformar mejor la dimensión características individuales. Por otro lado, es cierto que sería costoso y complejo obtener información para conformar las dimensiones de integración con el medio universitario, integración académica e integración social, pero la información completa para la dimensión características individuales puede obtenerse de manera natural mediante el cuestionario socioeconómico que se aplica a los estudiantes al ingreso. De hecho, también pueden ser incluidas preguntas que de algún modo midan la dimensión compromiso con la meta, que es la otra dimensión que Tinto establece

Es necesario llevar a cabo otras investigaciones en las que primero, para poder hablar de deserción, como era el propósito original, se tenga información confiable de una cohorte durante 10 años. De este modo es posible saber cuándo un alumno es desertor. Además, para enriquecer los resultados, podría obtenerse más información referida a las tres dimensiones que Vincent Tinto maneja. Los resultados en la Regresión de Cox quizás cambiarían ya que, al ser un método multivariado, desconocemos el efecto de las variables omitidas.

Este estudio solamente abarcó la División de Ciencias Básicas e Ingeniería de la universidad Autónoma Metropolitana, Unidad Iztapalapa, resultaría interesante para la universidad ampliar este estudio a las otras divisiones de la unidad Iztapalapa y a las otras Unidades que integran la universidad Autónoma Metropolitana. Esto permitiría comparar y establecer un resultado global para el conjunto de la institución. Además se sugiere a la División de Ciencias Básicas e Ingeniería de la universidad Autónoma Metropolitana, Unidad Iztapalapa, alargar el periodo regular de las Licenciaturas Es decir, que sea mayor de 12 trimestres, dado que se observó una tasa de rezago del 100%. Además, se recomienda que el proceso de selección se oriente fundamentalmente a seleccionar alumnos con porcentajes de aciertos altos en las áreas de razonamiento matemático y de Conocimientos Generales en el examen de admisión.

A su vez se recomienda, en principio, el Análisis de sobrevivencia como una técnica útil para estos estudios y, en segundo plano, el análisis de conglomerados como técnica complementaria.

CAPÍTULO 5
CONCLUSIONES

El problema de la deserción escolar es muy complejo. Por ello, no puede ser comprendido cabalmente por técnicas descriptivas o correlacionales. Son necesarias herramientas de análisis multivariado que permitan identificar, de manera oportuna, a los estudiantes con alto riesgo de abandono de los estudios, con el objeto de que las instituciones puedan tomar las decisiones adecuadas para propiciar su permanencia y conclusión de los estudios. Por ejemplo, el análisis de conglomerados permite formar grupos homogéneos y modelar su desempeño académico y el análisis de sobrevivencia puede ser utilizado para estimar la probabilidad, en función del tiempo, de retención de los estudiantes en el nivel universitario.

Dado que, en general, entre los estudiantes del nivel universitario existe un importante rezago escolar, conviene que los estudios sobre deserción, retención y abandono consideren el plazo máximo que los estudiantes pueden permanecer en una institución, para que los abandonos temporales no produzcan sesgos en los análisis.

El modelo de análisis y diagnóstico desarrollado en este trabajo permite estimar la probabilidad, en función del tiempo, de que el estudiante, dadas sus características, sea retenido en la universidad.

Al modelar el desempeño académico por medio del análisis de Conglomerados se identificaron como variables asociadas con un mayor potencial de éxito, al rendimiento en las áreas de razonamiento matemático y conocimientos generales. El área de razonamiento verbal no parece asociarse al mayor potencial de éxito.

Por otro lado, al analizar las funciones de sobrevivencia para cada variable en cada una de sus categorías se obtuvo que:

- Las mujeres tienen mayor probabilidad de sobrevivir que los hombres.

- La edad al ingreso que determina una probabilidad más alta de ser retenido en la universidad es de alrededor de 22 años.

- Los estudiantes que obtuvieron su certificado de educación media superior en escuelas incorporadas a la UNAM tienen mayor probabilidad de permanecer en la Institución que los estudiantes egresados del Colegio de Bachilleres, de escuelas incorporadas a la SEP o de escuelas clasificadas como "Otras".

- En relación con el examen de admisión, la probabilidad más alta de sobrevivencia al estado crítico no - retención fue para quienes obtuvieron al menos el 75% de aciertos en el área de razonamiento matemático y más del 50% en el área de conocimientos generales

- Las carreras cuyos estudiantes tienen mayor probabilidad de sobrevivencia, en la División de Ciencias Básicas e Ingeniería de la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, son Ingeniería Electrónica e Ingeniería en Computación y las que en este estudio resultaron con menor probabilidad de retener a sus estudiantes en la institución fueron Química, Física, Ingeniería Hidrológica e Ingeniería en Energía.

- El alcance de este estudio es muy limitado por varias razones:

  1) Sólo se estudió una cohorte, correspondiente a los dos períodos de ingreso de la universidad en un año lectivo.
  2) Sólo se manejó la información correspondiente a los 12 trimestres que constituyen el plazo normal establecido en la reglamentación universitaria para cursar las carreras analizadas
  3) Al ser una División con una tasa de rezago muy alta, no se pudieron evitar los sesgos de los abandonos temporales ni determinar si realmente los estudiantes con un alto potencial de éxito terminaron sus estudios.
  4) Las variables consideradas en el estudio sólo comprenden dos de las siete dimensiones planteadas por Vincent Tinto como necesarias para el estudio de la deserción: características individuales y antecedentes educativos. Incluso, la dimensión de características individuales se abordó en forma parcial, ya que no se contó con información relativa a características socioeconómicas de los estudiantes.

- Sería importante llevar a cabo más estudios integrando al menos las variables relacionadas al nivel socioeconómico que complementarían la dimensión características individuales. Asimismo, se estima conveniente que dichas investigaciones abarquen varias cohortes (dado que su comportamiento pudiera ser distinto) y que incluyan a las demás Divisiones de la Unidad Iztapalapa y de las otras dos unidades que integran a la Universidad Autónoma Metropolitana: Unidad Azcapotzalco y Unidad Xochimilco. Esto permitiría a esta institución contar con información valiosísima para la toma de decisiones orientadas a mejorar la permanencia de sus estudiantes y a elevar la eficiencia terminal.

BIBLIOGRAFÍA

- Altamira. Rodriguez (1997). El análisis de las trayectorias escolares como herramienta de evaluación de la actividad académica universitaria

- Bisquerra Alzina Rafael (1989). Introducción Conceptual al Análisis Multivariable Un enfoque informático con los paquetes SPSS-X, BMDP, LISREL y SPAD. Editorial PPU Barcelona.

- Braxton, John, *et al* (2000). Tinto´s separation stage and its influence on first-semester college student persistence . En: *Research in Higher education.* Volume 41, No. 2.

- Díaz de Cosío, Roger (1998). "Los desafíos de la educación superior mexicana" . En: Revista de la Educación Superior, No. 106. ANUIES.

- Chain, Ragueb (1995). Estudiantes universitarios: Trayectorias escolares Xalapa, UV-UAA. 329 pp.

- Dallas E., Johnson (2000). Métodos Multivariados Aplicados al Análisis de Datos. International Thompson Editores. 551pp.

- Lee, Elisa.T. (1992). Statistical Methods for Survival Data Analysis. Wiley Series in Probability and Mathematical Statistics.483pp.

- Martínez Arias, Rosario (1999) El Análisis Multivariante en la Investigación Científica. Editorial La Muralla. Madrid, España 143pp.

- Martínez Rizo, Felipe (1989). Diseño de investigación para el estudio de la deserción Enfoque cuantitativo transversal En: Trayectorias escolar en la educación superior COMPES-ANUIES, México.

- Mendez Ramírez, Ignacio *et. al.* (1984). El protocolo de la investigación. Lineamientos para su elaboración y análisis. Editorial Trillas. México. 210pp

- Miller, Rupert G (1981). Survival analysis / Rupert g. miller ; notes by gail gong ; problem solutions by Alvaro Muñoz. New york. Wiley. 238pp

- Moreu Jalón (1999) Estadística Informatizada Editorial Paraninfo Madrid, España. 140 pp.

- Muñiz Martelón, Patricia (1997). Trayectorias educativas y deserción universitaria en los ochenta. En: Temas de Hoy En la Educación superior. Núm. 19, ANUIES 190 pp.

- Murtaugh, Paul A, Leslie D. Burns y Jill Schusler (1999). Predicting the retention of university students. En: Research in Higher Education, Vol. 40, N° 3, pp 355-371.

- Perez López, Cesar (1996) Econometría y análisis estadístico multivariable con STATGRAPHICS. Técnicas avanzadas Madrid. 743pp.

- Rivas L. J y López H J (2000). Análisis de supervivencia. Cuadernos de Estadística. Ed La Muralla y Ed. Hespéndes.

- SPSS (1997). Advanced Statistical Analysis Using SPSS

- Tinto, Vincent (1992). El abandono de los estudios superiores. Una nueva perspectiva de las causas del abandono y su tratamiento. México, UNAM-ANUIES. 268 pp.

- Tinto, Vincent, *et al.* (1989). Trayectoria escolar en la educación superior. México, SEP-ANUIES 299 pp.

- Van Gennep, A. (1960). The Rites of passage, trans. By M. Viedon and G. Caffee. Chicago University: University of Chicago Press.

- Zubizarreta Armando F (1988). La aventura del trabajo intelectual México, Adisson Wesley. Colombia. 184pp.

GLOSARIO

Alumno desertor – nivel superior
Alumno de nivel superior que comunica a la administración de la institución educativa su abandono de los estudios, o que durante dos años sucesivos no realiza ninguna inscripción, o bien no acredita ningún curso.

Esta definición abarca tanto a los llamados *"desertores de derecho"* como a los *"desertores de hecho"*.

Alumno regular.
Alumno que se inscribe en un periodo educativo habiendo acreditado todos los requisitos del mismo.

Alumno rezagado
Alumno que se atrasa en las inscripciones que corresponden al trayecto escolar de su cohorte o en el egreso de la misma.
Esta definición permite análisis transversales y longitudinales de esta clase de población de alumnos. Pueden desagregarse los datos, de manera que se identifiquen alumnos cuyo rezago es permanente y otros cuyo rezago es provisional. Por otra parte, también se pueden identificar los momentos en los que haya mayor proporción de rezago, etcétera.

Análisis de Sobrevivencia:
Dada una variable de interés cuyos valores son registrados durante el tiempo, en un conjunto de individuos, hasta que ocurre un determinado evento de interés que esta en función de la variable medida, el objetivo del análisis de sobrevivencia para este tipo de información registrada es estimar, en función del tiempo, la probabilidad de que ocurra dicho evento de interés.

Carrera:
Conjunto de estudios y actividades que debe cursar y realizar un estudiante para obtener un título profesional.

Cohorte
Grupo de personas que inician sus estudios al mismo tiempo

La cohorte es la unidad fundamental del análisis estadístico, porque con base en ella se pueden agrupar y desagregar los datos referentes a los alumnos

Curva de Sobrevivencia:
Representación gráfica de la función de sobrevivencia

Desempeño académico:
Porcentaje de créditos aprobados por unidad de tiempo La unidad de tiempo en este trabajo es el trimestre lectivo.

Deserción:
Incluye cuatro posibilidades·

1. Es el abandono o suspensión voluntaria y definitiva de los estudios por parte del alumno, lo cual puede deberse a problemas tanto sociales como personales.
2. Por deficiencia académica, es la expulsión de alumnos de bajo rendimiento escolar
3. Por cambio de carrera (continúa el alumno en la misma institución pero pasa a pertenecer a otra cohorte)
4. Por expulsión disciplinada, la que se aplica a los alumnos que alteran el orden y la disciplina, quienes reciban esta sanción no pueden ingresar a ninguna escuela o facultad de la universidad.

Deserción:
Se define por el abandono que hace el alumno de los cursos o carreras a las que se ha inscrito, dejando de asistir a las clases y de cumplir las obligaciones fijadas, lo cual afecta la eficiencia terminal del conjunto Es un indicador que, tomando en cuenta el total de la deserciones de los alumnos, aprecia el comportamiento del flujo escolar de una generación.

Educación:
Medio fundamental y proceso permanente par ala adquisición, transmisión y acrecentamiento de los conocimientos y la cultura, que contribuye al desarrollo del individuo y la sociedad.

Eficiencia terminal.
Relación cuantitativa entre los alumnos que ingresan y los que egresan de una cohorte.
Esta relación cuantitativa se suele expresar porcentualmente Se le considera un índice de eficiencia interna institucional. Un error muy común en relación a la obtención de la eficiencia terminal es el considerar las cifras de *egresados en general* y no las de *egresados de una cohorte.* En el primer caso se incluyen los alumnos rezagados y los alumnos avanzados

Egresado
Alumno que ha cumplido con todos los requisitos académicos y administrativos correspondientes a un plan de estudios

Función de tiempo de sobrevivencia:
Es la caracterización de la distribución de los tiempos de sobrevivencia.

Función de sobrevivencia.

Se define como la probabilidad de que un individuo sobreviva más allá del tiempo *t.*

Función de densidad de probabilidad.

La función de densidad de probabilidad para el tiempo de sobrevivencia está definida como el límite de la probabilidad de que un individuo fallezca en el intervalo corto de tiempo (t, t+ Δ t) o simplemente la probabilidad de no sobrevivir en un pequeño intervalo por unidad de tiempo.

Función Hazard (Hazard Function).

Es la probabilidad de fallecimiento al comienzo del intervalo, o como el límite de la probabilidad de que un individuo fallezca en un pequeño intervalo de tiempo, (t,t+Δt) dado que el individuo ha sobrevivido al tiempo t.

Observaciones censuradas.
El análisis de sobrevivencia considera los casos en que el evento de interés no ha ocurrido al final del estudio o en un tiempo de análisis determinado. El tiempo exacto de sobrevivencia de estos sujetos se desconoce. Estas reciben el nombre de observaciones censuradas o tiempos censurados. También puede ocurrir este tipo de observación cuando no es posible por razones desconocidas seguir monitoreando a los individuos sujetos de estudio.

Método no paramétrico producto-límite de Kaplan y Meier.
Método no paramétrico que estima la función e sobrevivencia considerando a las observaciones censuradas.

Modelo de funciones hazard proporcionales de Cox.
La Regresión de Cox es un método semiparamétrico que permite estudiar datos de sobrevivencia. 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.

Retención:
Conjunto de integrantes de integrantes de una cohorte que permanecen inscritos a través de varios periodos escolares del nivel o niveles estudiados en el seguimiento del mismo.

a. Es el número de alumnos de una promoción que continúa sus estudios en el mismo grado o en el siguiente.
b. Es el número de alumnos que habiéndose matriculado en una año y grado o curso dado aparecen matriculados en el siguiente.

Rezago:
Consiste en el atraso en la inscripción a las asignaturas que según la secuencia que indica el plan de estudios, corresponden a la cohorte.

Tiempo de sobrevivencia:
Tiempo en el que ocurre el evento de interés

Trayectoria escolar
Se refiere a la cuantificación del comportamiento escolar de un conjunto de estudiantes (cohorte) durante su trayecto estancia educativa o establecimiento escolar, desde el ingreso, permanencia y egreso, hasta la conclusión de los créditos y requisitos académicos-administrativos que define el plan de estudios

- **Fuentes consultadas para la integración del glosario:**

Altamira Rodríguez, *El Análisis de las trayectorias escolares como herramienta de evaluación de la actividad académica universitaria*. 1997.

Lee T. Elisa T, Statistical Methods for Survival Data Analysis. Wisley Series in Probability and Mathematic Statistics.

Huerta Ibarra José y Allende Carlos María, "Aportación metodológica a la definición de las clases de alumnos". En Trayectoria escolar en la Educación Superior ANUIES México 1989.

Rangel Guerra Alfonso, Glosario de la Educación Superior. ANUIES. México 1988