UNIVERSIDAD NACIONAL DEL ALTIPLANO - PUNO

FACULTAD DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

"MODELO DE REGRESIÓN LOGÍSTICA CON DATOS CENSURADOS Y NO CENSURADOS PARA EL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 – II"

TESIS

PRESENTADA POR

Bach. LUISA PAOLA VALERIA ACURIO CRUZ

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO ESTADÍSTICO E INFORMÁTICO

PUNO – PERÚ

2014

UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO
FACULTAD DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

TESIS

"MODELO DE REGRESIÓN LOGÍSTICA CON DATOS CENSURADOS Y NO CENSURADOS PARA EL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 – II"

PRESENTADA POR

Bach. LUISA PAOLA VALERIA ACURIO CRUZ

A la Coordinación de Investigación de la Facultad de Ingeniería Estadística e Informática de la Universidad Nacional del Altiplano – Puno, para optar el Título Profesional de:

INGENIERO ESTADÍSTICO E INFORMÁTICO

APROBADA POR:

PRESIDENTE : _____
M.Sc. Edgar Eloy Carpio Vargas

PRIMER MIEMBRO : _____
M.Sc. Godofredo Quispe Mamani

SEGUNDO MIEMBRO : _____
M.Sc. Percy Huata Panca

DIRECTOR : _____
Mg. Emma Orfelinda Azañero de Aguirre

ASESOR : _____
Dr. Vladimiro Ibañez Quispe

2014

ÁREA: Estadística
TEMA: Modelos probabilísticos

DEDICATORIAS

A mis padres, por su infinito amor, esfuerzo y comprensión.

A mis hermanos, por mostrarme lo divertida, problemática y entendible que llega a ser la vida.

A los docentes de la Facultad de Ingeniería Estadística e Informática, por apoyarme en mi formación profesional y humanística.

AGRADECIMIENTOS

Quiero agradecer a mi familia, por comprenderme, apoyarme y motivarme en cada instante de mi vida.

A mi directora y asesor de la tesis, por guiarme a lo largo de esta aventura llamada investigación.

A el Dr. Luis Escobar, quien a través de la distancia me mostró lo maravilloso que significa la Estadística.

A los docentes de la facultad, por cuyas enseñanzas pude entender que cada día se puede mejorar.

A todos los Decanos, Directores de Escuelas, Coordinadores de Investigación y Docentes de las 35 escuelas profesionales de la universidad, quienes ayudaron con la recolección de información.

A mis compañeros de la Escuela profesional, por todos los momentos inolvidables.

Y por último a todas aquellas personas que han contribuido con esta tesis.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

RESUMEN

Relacionar la presencia del síndrome de Burnout con variables demográficas y socioeconómicas requiere de la estimación de modelos de regresión logística, la inclusión de datos censurados puede mejorar el modelo estimado, motivo por el cual en la presente investigación se tiene como objetivo comparar los modelos de regresión logística con datos censurados y no censurados para estimar la presencia del Síndrome de Burnout en los estudiantes de la Universidad Nacional del Altiplano Puno, 2013 – II.

En la presente investigación se trabajó con una muestra estratificada por escuela profesional de 511 estudiantes. Las variables a analizar son la presencia del síndrome de Burnout como variable independiente y sus factores, la medición de los factores se realizó mediante un cuestionario preparado para la investigación, y para la medición del Síndrome de Burnout se utilizó el test Maslach Burnout Inventory. En la comparación de los modelos se utilizó las medidas del sesgo asintótico, error medio integrado y la pseudoverosimilitud, de las funciones de densidad de los modelos de regresión logística.

Se determinó que el 13.11% de los estudiantes de la Universidad Nacional del Altiplano Puno tiene el síndrome de Burnout. El mejor modelo para estimar la presencia del síndrome de Burnout fue el modelo de regresión logística con datos censurados, cuya ecuación fue: $p_i = \frac{1}{1+\exp(-1.749x_{3i}-1.560x_{4i}-0.276x_{5i}-1.528x_{8i})}$, el cual mostro un menor sesgo asintótico y un error medio integrado, una mayor pseudoverosimilitud que el modelo de regresión logística con datos no censurados, cuya ecuación fue: $p_i = \frac{1}{1+\exp(-3.005x_{3i}-2.099x_{4i}-0.177x_{5i}-2.318x_{8i})}$.

PALABRAS CLAVE: Regresión, Datos Censurados, Burnout, Estudiantes

ABSTRACT

Relate the presence of burnout syndrome with demographic and socioeconomic variables requires estimation of logistic regression models, the inclusion of censored data can improve the estimated model, which is why in this research aims to compare the regression models logistics censored and uncensored data to estimate the presence of burnout syndrome in the students of the National University of Altiplano Puno, 2013 - II.

The variables analyzed are the presence of burnout syndrome as an independent variable and its factors, measurement of the factors was performed using a questionnaire prepared for research, and for measuring burnout syndrome test was used Maslach Burnout Inventory . In comparing the models measures the asymptotic bias was used, integrated and pseudoverosimilitud mean error of the density functions of the logistic regression models.

It was found that 13.11 % of the students at the National University of Altiplano Puno have Burnout Syndrome. The best model for estimating the presence of burnout syndrome was the logistic regression model with censored data, the equation was: $p_i = \frac{1}{1+\exp(-1.749x_{3i}-1.560x_{4i}-0.276x_{5i}-1.528x_{8i})}$. That showed a lower asymptotic bias and a mean error integrated, a pseudoverosimilitud greater the logistic regression model with uncensored data, the equation was: $p_i = \frac{1}{1+\exp(-3.005x_{3i}-2.099x_{4i}-0.177x_{5i}-2.318x_{8i})}$.

KEYWORDS: Regression, Censored Data, Burnout, Students

INTRODUCCIÓN

Las técnicas desarrolladas a lo largo del tiempo para la estadística nos indican la mejor forma de recoger, procesar, analizar e interpretar la información. La que es de suma importancia para el área en el que investiga, por lo que perder parte de la información, es perder tiempo y dinero. Por tal motivo, en los últimos años se viene desarrollando técnicas para aprovechar al máximo toda la información de la que se dispone (minería de datos y análisis de supervivencia de datos), los datos censurados eran considerados información perdida, por tal motivo no se los incluía en modelos de regresión logística, es este problema el que lleva a los estadísticos a investigar nuevas formas de trabajar con estos datos e implícitamente desarrollar nuevas técnicas que permitan el aprovechamiento de esta información.

Una de esas técnicas es la estimación de modelos de regresión no paramétrica, mediante la cual podemos estimar probabilidades de ocurrencia de sucesos de interés, esta técnica puede utilizas los llamados modelos de regresión polinómica y los modelos tipo kernel, estos últimos fueron los utilizados en la presente investigación; se utilizó la función de núcleo de Epanechnikov, puesto que es la más utilizada para el caso de estimación de funciones de regresión logística. Para dicho núcleo se calculó tres diferentes parámetros de suavizados. Del que se seleccionó el parámetro Plug- in (Polansky y Baker) puesto que mostró menor error frente a los demás parámetros, en base a este estimador se obtuvo la función de distribución y las funciones de densidad para los datos mediante el software R project.

Para realizar la comparación de ambos modelos se utilizó, las medidas como la pseudoverosimilitud (medida de máxima verosimilitud obtenida para los modelos estimados en regresión no paramétrica), el error medio integrado, y el sesgo asintótico.

La variable principal la presencia del Síndrome de Burnout puede llegar a tener consecuencias graves en la salud, produciendo otras enfermedades de consecuencias no solo psicológicas sino también físicas, las universidades preocupadas por ofrecer enseñanza de calidad deben tener en cuenta las variables relacionadas con el proceso de enseñanza – aprendizaje, como la falta de adaptación al sistema educativo que sufren los estudiantes de semestres inferiores, por lo que un conocimiento del nivel de estrés en los estudiantes universitarios y sus factores más influyentes se hace necesario.

La presente investigación se desarrolló con la finalidad principal de comparar el modelo de regresión logística con datos censurados y no censurados para el Síndrome de Burnout de los estudiantes de la Universidad Nacional del Altiplano, 2013 – II.

La estructura de la presente investigación es la siguiente:

En el Capítulo I se realizó la identificación del problema, planteamiento de objetivos e hipótesis de investigación.

En el Capítulo II se desarrolló la búsqueda de la información disponible que nos ayudó en la resolución del problema, la teoría disponible de los métodos estadísticos que se utilizaran para resolver el problema.

En el Capítulo III se observó la metodología utilizada para la resolución del problema: recolección de datos, procesamiento de datos, obtención de resultados según los objetivos planteados.

En el Capítulo IV se realizó el análisis de datos y obtención de resultados, la contrastación de nuestros resultados con los resultados obtenidos por otros investigadores.

En el Capitulo V se explican las principales conclusiones de la investigación.

En el Capítulo VI se dan las recomendaciones y sugerencias para próximas investigaciones.

Por último se presentan la Bibliografía y los Anexos de la investigación.

CAPÍTULO I

PLAN DE INVESTIGACIÓN

1.1 EL PROBLEMA

1.1.1 Definición del problema

La determinación de un modelo estadístico para cualquier tipo de variables necesita la recopi0lación de información, cuando se aplican instrumentos de medición en los que se requiere seguir la evolución de la muestra para medir el grado de estrés, este tiempo de seguimiento puede verse turbado, es aquí cuando se produce la censura de datos. Cuando esto sucede nos vemos en la necesidad de excluir dichos datos para la estimación de modelos de regresión logística.

Dado que los datos censurados no eran utilizados para realizar una inferencia estadística, se producía una gran pérdida de información, hasta hace algunos años se cuestionaba si su inclusión en el modelamiento puede contribuir o no a mejorar la fiabilidad de los modelos de regresión logística, y de esta forma obtener un modelo con indicadores más confiables. Por ello es que probaremos si la inclusión de los datos censurados en el

modelo de regresión logística para el Síndrome de Burnout de los estudiantes de la Universidad Nacional del Altiplano Puno, produce mejores estimadores.

El estrés es una enfermedad que ha acompañado a la humanidad desde el principio de los tiempos, sin embargo en los últimos años se ha notado que gran parte de la población es afectada de manera negativa por el estrés más específicamente por el Síndrome de Burnout, la Universidad Nacional del Altiplano no es ajena a esta realidad, los estudiantes son sometidos a presión constante a lo largo del desarrollo de su carrera universitaria lo que produce estrés en ellos, esto nos inspira para lograr tener una visión más sólida de este problema, e investigar cuales son las variables que más afectan a la presencia del síndrome. Por lo que estadísticamente esto sería resuelto con un modelo de regresión logístico clásico, sin embargo si los datos fueran censurados, tendríamos que descartar una parte de la información obtenida. Pero como se mencionó antes es posible realizar un modelo que incluya datos censurados, esto nos conduce a preguntar:

1.1.2 Formulación del problema

¿Existe diferencia significativa entre el modelo de regresión logística que incluye datos censurados y el modelo de regresión logística con datos no censurados para estimar la presencia del

Síndrome de Burnout en los estudiantes de la Universidad Nacional del Altiplano Puno, 2013 - II?

1.2 OBJETIVOS

1.2.1 Objetivo general

Determinar el mejor modelo de regresión logística para la estimación de la presencia del Síndrome de Burnout en los estudiantes de la Universidad Nacional del Altiplano Puno, 2013 – II.

1.2.2 Objetivos específicos

- Determinar el nivel de presencia del Síndrome de Burnout en los estudiantes de la Universidad Nacional del Altiplano Puno 2013 - II.
- Determinar el modelo de regresión logística que contemple los datos censurados para la estimación de la presencia el Síndrome de Burnout en los estudiantes de la Universidad Nacional del Altiplano Puno 2013 – II.
- Determinar el modelo de regresión logística que excluya los datos censurados para la estimación de la presencia el Síndrome de Burnout en los estudiantes de la Universidad Nacional del Altiplano Puno 2013 – II.
- Comparar los modelos de regresión logística con datos censurados y no censurados para la estimación de la

presencia del Síndrome de Burnout en los estudiantes de la Universidad Nacional del Altiplano Puno, 2013 – II.

1.3 HIPÓTESIS

1.3.1 Hipótesis general

El modelo de regresión logística con datos censurados tiene mejores indicadores que el modelo de regresión logística con datos no censurados para la estimación de la presencia del Síndrome de Burnout en los estudiantes de la Universidad Nacional del Altiplano Puno, 2013 – II.

CAPÍTULO II

MARCO TEÓRICO

2.1 ANTECEDENTES DE LA INVESTIGACIÓN

CASTILLO E., LÓPEZ M., RAMOS A., FERNÁNDEZ A. (2006). *Influencia del número de datos censurados en la evaluación del campo S-N.* (Tesis de pregrado). Universidad de Cantabria. El objetivo principal de este trabajo de investigación fue el de determinar el número máximo de ensayos censurados que cabría admitir en un programa experimental a fin de que no influyan negativamente en el ajuste. Concluyendo en que la consideración de los datos censurados en el ajuste de resultados de fatiga para la determinación del campo S-N, supone una potencial mejora en la fiabilidad del ajuste de parámetros.

FERNANDEZ, E. (2009). Estrés percibido, estrategias de afrontamiento y sentido de coherencia en estudiantes de enfermería: su asociación con salud psicológica y estabilidad emocional. (Tesis doctoral). Universidad de León. Cuyo objetivo fue el de probar el grado de relación que existe entre las variables como el sentido de coherencia, las estrategias de afrontamiento, la salud percibida, el cansancio emocional, la autoestima y la satisfacción con los estudios. Concluyendo en que existe una relación entre el estrés percibido

reciente y el cansancio emocional, además aclara que ambas variables explican el 61% de la varianza total en el modelo de predicción.

JACOME, A., (2005). *Estimación pre suavizada de las funciones de densidad y distribución con datos censurados.* (Tesis doctoral). Universidad de Coruña. Hizo una remembranza y explicación muy detallada de los datos censurados y modelos adecuados para esta clase de datos, lo cual la condujo a concluir en que la inclusión de los datos censurados en el modelamiento, teoricamente nos conlleva a cometer un error mucho menor que el normal al momento de predecir.

LÓPEZ C., ZEGARRA A., CUBA V., (2006). *Factores asociados al Síndrome de Burnout en enfermeras de emergencia del Hospital Nacional Guillermo Almenara Irigoyen*. (Tesis de pregrado). Universidad Peruana Unión. El objetivo principal de esta tesis fue determinar los factores asociados al síndrome de Burnout, además de determinar la asociación entre factores profesionales y laborales asociados al síndrome de Burnout. Concluyendo en que existe una baja asociación estadística (p=0.005) entre los factores profesionales conflictiva y ambigüedad de rol, también baja conciliación estadística entre los factores laborales (sobrecarga laboral y supervisión) y el síndrome de Burnout.

2.2 BASE TEÓRICA

2.2.1 Estimación de modelos de regresión no paramétrico

Formulación de un modelos no paramétrico

Consideremos un modelo de regresión múltiple no paramétrico habitualmente formulado como:

$$Y_i = m(X_{i1}, \dots, X_{id}) + \varepsilon_i$$

Donde $\{X_i, Y_i, i = 1, \dots, n\}$ son observaciones independiente e idénticamente distribuidas (i.i.d.) de una variable p+1 dimensional (X, Y) y los residuos $\{\varepsilon_i, i = 1, \dots, n\}$ son variables aleatorias independientes son $E[\varepsilon_i] = 0$, $Var[\varepsilon_i] = \sigma^2$ para todo $i = 0, \dots, n$. Aquí m es la funcion de regresión y modeliza la relación de dependencia entre la variable explicativa X y la variable respuesta Y.

El objetivo en el problema de regresión es la estimación de la función de regresión m que suponemos desconocida. En dicha tarea la formulación clásica consiste en asumir que m pertenece a alguna clase paramétrica de funciones, con lo cual el problema se resolvería encontrando aquella función dentro de la clase que sea óptima bajo algún criterio de optimalidad (habitualmente el criterio de mínimos cuadrados). La solución paramétrica asume solamente que la función es “suave”, entendiendo por esta suavidad en términos de derivabilidad.

Para la estimación no paramétrica de la función m, que bajo la formulación inicial se trata de una función $p - valuada$, existen diversos métodos en la literatura en las últimas décadas. Entre ellos destacamos los métodos basados en núcleos o *Kernels*, como son el estimador de Nadaraya-Watson y en general los métodos basados en estimación polinomial local y los métodos basados en *splines,* Algunos otros enfoques como el gráfico de dispersión localmente ponderado suavizado (*locally weighted scatter plot smoothig* (LOWESS)), métodos basados en wavelet y otros enfoques basados en series ortogonales también son de uso frecuente. Los primeros se conocen como métodos de suavizado (*smoothing*) e involucran en su definición la denominada función núcleo o *kernel*, denotada por $K(t)$, y habitualmente definida como un densidad simétrica y con soporte compacto (PARRA MURCIEGO, 2011), y un parámetro positivo, denotado por h, que define el tamaño de los entornos locales definidos alrededor de cada punto de estimación.

El modelo aditivo no paramétrico

Se dice que una función de regresión m satisface un modelo aditivo si tiene la forma

$$Y_i = a + \sum_{j=1}^{d} m_j(X_{ij}) + \varepsilon_i$$

Con $E[\varepsilon_i] = 0, Var[\varepsilon_i] = \sigma^2$ para todo $i = 1, \dots, n$, y además para asegurar la identificabilidad de las componentes aditivas m_j suponemos que $E[m_j(X_{ij})] = 0$ para todo $j = 1, \dots, d$.

Obsérvese que asumir un modelo aditivo supone que las covariables tienen efectos separados en la variable respuesta y vienen presentados por las funciones m_j, denominadas componentes del modelos aditivo y que suponemos desconocidas.

Bajo la aproximación anterior de nuevo es posible resolver el problema mediante la aproximación paramétrica clásica bien estimar las funciones m_j no paramétricamente. Obsérvese sin embargo que en cierto modo el modelo aditivo está a medio camino entre el modelo de regresión lineal múltiple paramétrico (que combina aditivamente transformaciones lineales de las variables, $m_j(x_{ij})$) y el modelo de regresión múltiple no paramétrico (PARRA MURCIEGO, 2011).

Estimación del modelo aditivo no paramétrico

Una vez definido el modelo, el paso siguiente consiste en estimarlo a partir de las n-observaciones disponibles. Obsérvese que $E[Y_i] = a$ (ya que $E[\varepsilon_i] = 0\ y\ E[m_j(X_{ij})] = 0$). Además, si el parámetro a y todas las funciones m_j fueran conocidos, excepto la función m_k, entonces esta podría estimarse mediante cualquier estimador no paramétrico univariante (por ejemplo,

mediante un ajuste lineal local). Bastaría con aplicar ese estimador al conjunto de datos (X_i, Y_i^k), donde

$$Y_i^k = Y_i + a + \sum_{j=1, j\neq k}^{d} m_j(X_{ij})$$

Esta observación llevará a proponer el algoritmo conocido como *Backfitting* y su posterior mejora, el algoritmo *Smooth Backfitting,* para estimar este tipo de modelos. Aunque existen muchos otros métodos, todos ellos parten de la idea del algoritmo *Backfitting* o del método de integración marginal de Linton y Nielsen.

Regresión polinomial local

Si suponemos que la función de regresión m tiene p derivadas en un punto x_0, entonces vía el teorema de Taylor tenemos una aproximación de este tipo para los valores en un entorno de x_0.

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m^n(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^p(x_0)}{p!}(x - x_0)^p$$

Esto justifica que se puede aproximar localmente m por funciones polinomicas de grado p.

$$P_p(x) = \sum_{j=0}^{p} B_j(x - x_0)^j$$

Así, se obtienen estimadores de los coeficientes $\widehat{B_j}$ con $j = 0, .., p$ y entonces, observando la expresión, vemos que la estimación del término independiente $\widehat{B_0}$ será un estimador de m en x_0 y el resto de coeficiente $\widehat{B_j}$ proporcionan estimaciones de sus derivadas.

Por eso, con el fin de estimar m localmente mediante polinomios de grado p consideraremos un problema de mínimos cuadrados ponderados:

$$min \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} B_j (X_i - x_0)^j \right\}^2 k_h(X_i - x_0)$$

Donde:

h : Es un parámetro denominado ancho de banda o parámetro de suavizado que controla las observaciones que caen en cada entorno.

$K_h(u) = h^{-1}K(\frac{u}{h})$, donde la función $K(.)$, que se denomina función núcleo. Dicha función define las ponderaciones que se asignara a cada observación en el entorno local considerado. Habitualmente se supone una densidad simétrica y con soporte compacto.

p : Es el grado del ajuste polinomial local.

Además, como casos particulares se puede obtener el conocido *estimador núcleo de Nadaraya-Watson,* que supone realizar ajustes polinomiales locales de *grado cero*, y también cuando el ajuste polinomial es de *grado uno*, se obtiene el denominado *estimador lineal local* (BOUKICHOU ABDELKADER, 2010). Si bien estos ajustes constantes han sido estudiados y ampliamente utilizados por teóricos y analistas de datos, es el ajuste lineal el que ha mostrado ser más conveniente y el más usado actualmente en la práctica (CLEVELAND, 1979).

Función de distribución (función tipo núcleo)

Dada un muestra de n observaciones reales $X_1, \dots, X_n$ definiremos la estimación tipo núcleo como:

$$\widehat{f_n}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \qquad 1.1$$

Donde $K(x)$ es la función denominada Kernel, función núcleo o función de peso, que satisface ciertas condiciones de regularidad, generalmente es una función de densidad simétrica como por ejemplo la de la distribución normal, y $\{h_n\}$ es una secuencia de constantes positivas conocidas como ancho de ventana, parámetro de suavización o *bandwith.*

El estimador núcleo puede interpretarse como una suma de protuberancias situadas en las observaciones (MIÑARRO, 1998). La función núcleo determina la forma de las protuberancias mientras que el parámetro h_n determina la anchura. Al igual que en el histograma h_n también determina la cantidad de suavización de la estimación, siendo el limite cuando h_n tiende a cero una suma de funciones delta de Dirac en los punto de la observaciones. También puede interpretarse como una transformación continua de la función de distribución empírica de acuerdo a la función $K(x)$ que se encarga de redistribuir la masa de probabilidad $\frac{1}{n}$ en rededor de cada punto muestral.

Un inconveniente en la estimación tipo núcleo como se observa en la Figura 1 es que al ser el parámetro de ventana fijo a lo largo de toda la muestra, existe la tendencia a presentarse distorsiones en las colas de la estimación.

TABLA 1

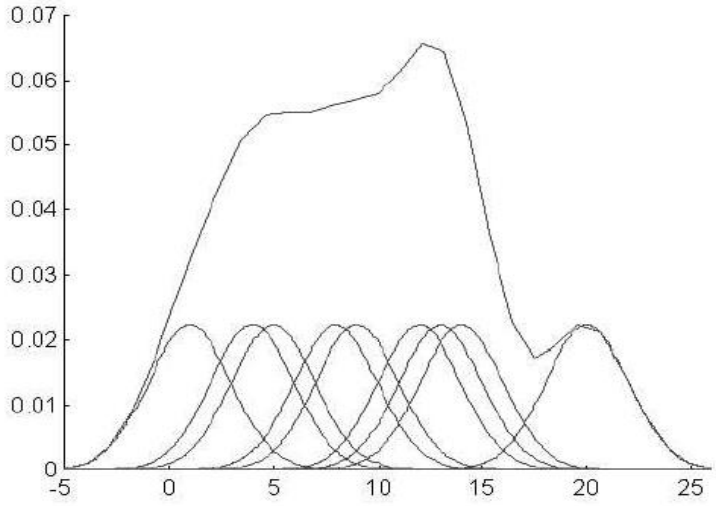
PRINCIPALES FUNCIONES NÚCLEO USADAS EN LA ESTIMACIÓN NO PARAMÉTRICA DE LA FUNCIÓN DE DENSIDAD

FUNCIÓN DE NÚCLEO	NOMBRE DE LA FUNCIÓN DE NÚCLEO
$\boldsymbol{K(u) = \frac{3}{4}(1-u^2)1_{\{\lvert u\rvert \le 1\}}}$	Núcleo de Epanechnikov
$\boldsymbol{K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}}$	Núcleo Gaussiano
$\boldsymbol{K(u) = (1-\lvert u\rvert)\, 1_{\{\lvert u\rvert \le 1\}}}$	Núcleo Triangular
$\boldsymbol{K(u) = \frac{15}{16}(1-u^2)^2 1_{\{\lvert u\rvert \le 1\}}}$	Núcleo Cuártico
$\boldsymbol{K(u) = \frac{1}{2} 1_{\{\lvert u\rvert \le 1\}}}$	Núcleo Uniforme

Fuente: Tesis Estimación pre suavizada de las funciones de densidad y distribución con datos censurados , Autor Amalia Jácome Pumar

FIGURA 1

ESTIMACIÓN TIPO NÚCLEO MOSTRANDO NÚCLEOS INDIVIDUALES



Fuente: Tabla 1

Propiedades del estimador tipo núcleo

- Consistencia

  Parzen (1962) Basándose en un teorema previo de Bochner (1955), que presentamos a continuación, estudia el sesgo y la consistencia en un punto x para las estimaciones tipo núcleo donde la función núcleo es una función simétrica y acotada que verifica

  $$\int_{-\infty}^{\infty}|K(x)|dx < \infty \qquad (1.2)$$

  $$\lim_{x\to\infty}|xK(x)| = 0 \qquad (1.3)$$

  $$\int_{-\infty}^{\infty}K(x)dx = 0 \qquad (1.4)$$

  Condiciones que son satisfechas por cualquiera de las funciones núcleo presentadas en la TABLA 1.

Teorema 4 (BOCHNER, 1955) Sea $K(y)$ una función Borel acotada que satisface la condiciones (1.2) y (1.3). Sea $g \in \mathcal{L}^1$.Sea

$$g_n(x) = \frac{1}{h_n}\int_{-\infty}^{\infty} K\left(\frac{y}{h}\right) g(x-y)dy \qquad (1.5)$$

Donde $\{h_n\}$ es una consecuencia de constantes positivas que satisfacen $\lim_{n\to\infty} h_n = 0$. Entonces si x es un punto de continuidad de g.

$$\lim_{n\to\infty} g_n(x) = g(x)\int_{-\infty}^{\infty} K(y)dy \qquad (1.6)$$

Demostración:

Notemos en primer lugar que

$$g_n(x) - g(x)\int_{-\infty}^{\infty} K(y)dy$$

$$= \int_{-\infty}^{\infty}\{g(x-y) - g(x)\}\frac{1}{h_n}K\left(\frac{y}{h}\right)dy$$

Sea ahora $\delta > 0$, y dividamos el dominio de integración en dos regiones, $|y| \leq \delta$ y $|y| > \delta$. Entonces

$$|g_n(x) - g(x)\int_{-\infty}^{\infty} K(y)dy| \leq \sup_{|y|\leq\delta}|g(x-y) - g(x)|\int_{|z|\leq\delta/h_n}|K(z)|dz +$$

$$\int_{|z|\leq\delta}\frac{|g(x-y)|}{y}\frac{y}{h_n}K\left(\frac{y}{h}\right)dy + |g(x)|\int_{|y|\geq\delta}\frac{1}{h_n}K\left(\frac{y}{h}\right)dy \leq \left|\leq \sup_{|y|\leq\delta}|g(x-y) - g(x)|\int_{-\infty}^{\infty}|K(z)|dz + \frac{1}{\delta}\right|\leq \sup_{|y|\geq\frac{\delta}{h_n}}|zK(z)|\int_{-\infty}^{\infty}|g(y)dy +$$

$$|g(x)|\int_{|z|\geq\delta/h_n}|K(z)dz|$$

Cuando $n \to \infty$, debido a que $h_n \to 0$, el segundo y tercer término tiende a cero, ya que $g \in \mathcal{L}^1$ y $\lim_{y\to\infty}|xK(x)| = 0$. Haciendo entonces que $\delta \to 0$, el primer termino tiende a

cero debido a que $K \in \mathcal{L}^1$ y a que x es un punto de continuidad de g.

Teniendo en cuenta que

$$E[\hat{f}_n(x)] = \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{h_n}K\left(\frac{x - X_i}{h_n}\right)\right]$$

$$= E\left[\frac{1}{h_n}K\left(\frac{x-y}{h}\right)\right]$$

$$= \int_{-\infty}^{\infty} \frac{1}{h_n}K\left(\frac{x-y}{h}\right) f(y)dy$$

Del teorema anterior se deduce lo siguiente

Corolario 1: La estimación definida en 1.1 es asintóticamente insesgada en todos los punto x en los cuales la función de densidad de probabilidad es continua si las constantes h_n satisfacen $\lim_{n\to\infty} h_n = 0$ y si la función $K(y)$ satisface (1.2) , (1.3) y (1.4).

Teorema 5 El estimador $\hat{f}_n(x)$ definido en (1.1) es consistente, es decir $MSE[\hat{f}_n(x)] \to 0\ \forall x \in \mathbb{R}$ cuando $n \to \infty$, si añadimos la condición adicional de que $\lim_{n\to\infty} nh_n = \infty$.

Demostración:

En efecto, tengamos en cuenta que

$$Var[\hat{f}_n(x)] = \frac{1}{n}Var\left[\frac{1}{h_n}K\left(\frac{x-y}{h}\right)\right] \qquad (1.9)$$

Además

$$\frac{1}{n}Var\left[\frac{1}{h_n}K\left(\frac{x-y}{h}\right)\right] \leq \frac{1}{n}E\left[\left(\frac{1}{h_n}K\left(\frac{x-y}{h}\right)\right)^2\right] \qquad (1.10)$$

$$\frac{1}{h_n n}\left[\frac{1}{h_n}\int_{-\infty}^{\infty}\left(K\left(\frac{x-y}{h}\right)\right)^2 f(y)dy\right] \qquad (1.11)$$

y por el Teorema 4

$$\frac{1}{h_n}\int_{-\infty}^{\infty}\left(K\left(\frac{x-y}{h}\right)\right)^2 f(y)dy \to f(x)\int_{-\infty}^{\infty}K^2(y)\,d(y) \qquad (1.12)$$

Ya que $\int_{-\infty}^{\infty}K^2(y)dy < \infty$.Es por tanto evidente que

$$\lim_{n\to\infty} Var[\hat{f}_n(x)] \to 0 \quad \text{si} \quad \lim_{n\to\infty} h_n n = \infty \qquad (1.13)$$

Finalmente al ser

$$MSE\left[\hat{f}_n(x)\right] = Var\left[\hat{f}_n(x)\right] + sesgo^2[\hat{f}_n(x)] \qquad (1.14)$$

Teniendo en cuenta Corolario 1 el Teorema queda demostrado.

Este resultado ilustra perfectamente el problema básico de la estimación no paramétrica. Una rápida convergencia al cero del parámetro h_n provoca una disminución del sesgo, pero sin embargo la varianza aumentaría de forma considerable. El ancho de ventana ideal debe converger a cero pero a un ritmo más lenta que n^{-1}.

Una axiomática más reciente para las funciones núcleo es la propuesta por (Nadaraya, 1989), donde diremos que una función $K(x)$ pertenece a la clase H_s ($s \geq 2$ es un número par) o bien que es un Kernel de orden s, si satisface las siguientes condiciones de regularidad

$$K(x) = K(-x) \qquad (1.15)$$

$$\int_{-\infty}^{\infty}K(x)dx = 1 \qquad (1.16)$$

$$\sup_{-\infty < x < \infty}|K(x)| < \infty \qquad (1.17)$$

$$\int_{-\infty}^{\infty} x^i K(x)dx = 0 \quad i = 1, \dots, s-1 \tag{1.18}$$

$$\int_{-\infty}^{\infty} x^s K(x)dx = k_s \neq 0 \tag{1.19}$$

$$\int_{-\infty}^{\infty} x^s K(x)dx < \infty \tag{1.20}$$

Destaquemos que si K es una función de densidad debe verificarse

$$k_2 = \int_{-\infty}^{\infty} x^2 K(x)dx > 0$$

Es posible hacer que $k_2 = 0$ si permitimos que K puede tomar valores negativos.

Si K es una función núcleo simétrica, entonces s debe ser un número par.

Si utilizamos como núcleo una función de densidad simétrica, supuesto bastante común y sobre el que se basa el siguiente apartado, estamos considerando un número de orden 2. En particular corresponden a este modelo las funciones presentadas en la TABLA 1.

- Minimización del AMISE

  La determinación del ancho de ventana se realiza de modo que se minimice algún tipo de error. En general se utiliza como medida del error cuadrático medio integrado. Y se minimiza una aproximación al mismo.

Podemos utilizar (1.8) para obtener una expresión aproximada del sesgo. Hagamos en (1.8) el cambio de variable $y = x - h_n t$, obtenemos

$$E\left[\hat{f}_n(x)\right] = \int_{-\infty}^{\infty} K(t) f(x - h_n t) dt \qquad (1.21)$$

Y haciendo un desarrollo de Taylor en el punto cero

$$f(x - h_n t) = f(x) - f'(x) h_n t + f''(x) \frac{h_n^2 t^2}{2} + \cdots$$

Y sustituyendo (1.21), obtenemos teniendo en cuenta (1.18)

$$E\left[\hat{f}_n(x)\right] = f(x) + \frac{h_n^2 f''(x) k_2}{2} + O(h^4) \qquad (1.22)$$

Por consiguiente el sesgo adopta la forma de

$$sesgo\left[\hat{f}_n(x)\right] = \frac{h_n^2 f''(x) k_2}{2} + O(h^4) \qquad (1.23)$$

A partir de (1.23) podemos escribir

$$AISB = \frac{1}{4} h_n^4 k_2^2 \int_{-\infty}^{\infty} f''(x)^2 dx \qquad (1.24)$$

Y de (1.9) se comprueba que

$$Var\left[\hat{f}_n(x)\right] = \frac{1}{n} \int_{-\infty}^{\infty} \frac{1}{h_n^2} \left(K\left(\frac{x - y}{h}\right) \right)^2 f(y) g y - \frac{1}{n} \{ f(x) + sesgo\left[\hat{f}_n(x)\right] \}^2$$

$$\approx \frac{1}{n h_n} \int_{-\infty}^{\infty} f(x - h_n t) K^2(t) dt - \frac{1}{n} \{ f(x) + O(h_n^2) \}^2 \qquad (1.25)$$

Usando la sustitución $y = x - h_n t$ y la aproximación (1.23) para el sesgo. Suponiendo un valor de h_n pequeño y un valor de n grande y expandiendo $f(x - h_n t)$ en serie de Taylor, obtenemos

$$Var\left[\hat{f}_n(x)\right] = \frac{1}{nh_n} f(x) \int_{-\infty}^{\infty} K^2(t)dt \qquad (1.26)$$

Integrando ahora tenemos

$$AIV = \frac{1}{nh_n} \int_{-\infty}^{\infty} K^2(t)dt \qquad (1.27)$$

Resultando finalmente a partir de (1.23) y (1.26)

$$AMISE\left[\hat{f}_n(x)\right] = \frac{1}{4} h_n^4 k_2^2 \int_{-\infty}^{\infty} f''(x)^2 dx + \frac{1}{nh_n} \int_{-\infty}^{\infty} K^2(t)d \qquad (1.28)$$

Busquemos ahora el valor de h_n que minimiza la expresión anterior, obtenemos

$$h_{opt} = \left\{ \frac{\int_{-\infty}^{\infty} K^2(t)dt}{\int_{-\infty}^{\infty} f''(x)^2 dx} \right\}^{1/5} n^{-1/5} k_2^{-2/5} \qquad (1.29)$$

Comprobamos como efectivamente la convergencia de h_n hacia cero es de orden $n^{-1/5}$, menor que n^{-1}. Debe notarse la dependencia del valor óptimo respecto a la densidad desconocida que se desea estimar, lo que impide que sea calculable directamente.

Substituyendo (1.29) en (1.28) obtenemos

$$AMISE^* = \frac{5}{4} \{K_2^2 R^4(K)\}^{1/5} \{R(f'')\}^{1/5} n^{-4/5} \qquad (1.30)$$

- Elección del parámetro de suavización.

Una de las posibilidades a la hora de elegir el parámetro de suavización óptimo es tomar como referencia una distribución estándar para obtener el valor de $\int_{-\infty}^{\infty} f''(x)^2 dx$ en la expresión (1.29). Una de las distribuciones más

utilizadas es la distribución normal de media cero y varianza σ^2, resultando entonces

$$\int_{-\infty}^{\infty} f''(x)^2 dx = \frac{3}{8}\pi^{-1/2}\sigma^{-5} \approx 0.21\sigma^{-5} \qquad (1.31)$$

Utilizando ahora la función núcleo de Gauss y sustituyendo (46) en (1.29) obtenemos el ancho de ventana (banda o parámetro de suavización) óptimo, resultando:

$$h_{opt} = (4\pi)^{-1/10}\left(\frac{3}{8}\pi^{-1/2}\right)^{-1/5}\sigma n^{-1/5} = 1.06\sigma n^{-1/5} \qquad (1.32)$$

Donde σ puede ser substituida por una estimación de la varianza a partir de los datos.

La utilización de (1.32) será adecuada si la población se asemeja en su distribución a la de la normal, sin embargo si trabajamos con poblaciones multimodales se produciría una sobresuavización de la estimación, (BOWMAN, A comparative study of some kernel-based nonparametric density estimators, 1985) (SILVERMAN, 1986). Una posible

modificación del parámetro de suavización es:

$$h_n = 1.06An^{-1/5} \qquad (1.33)$$

Donde $A = \min(desviacion\ estandar, rango\ intercuartil/1.349)$, comprobando que se comporta bien trabajando con densidades unimodales moderadamente bimodales.

Silverman también sugiere la reducción del factor 1.06 en (1.33); y propone un nuevo valor del parámetro h_n

$$h_n = 0.9\,An^{-1/5} \qquad (1.34)$$

- Selección de la función núcleo óptima

En (1.30) denominaremos

$$C(K) = \{k_2^2 R^4(K)\}^{1/5} \qquad (2.35)$$

En primer lugar veamos que $C(K)$ es invariante frente a transformaciones de la forma

$$K_\delta(.) = \frac{1}{\delta} K\left(\frac{\cdot}{\delta}\right) \quad \delta > 0$$

En efecto

$$\int K_\delta(x)x^2dx = \frac{1}{\delta}\int k(\frac{x}{\delta})x^2dx = \int K(u)\delta^2u^2du = \delta^2k_2$$

Además

$$R(K_\delta) = \int K_\delta^2(x)dx = \frac{1}{\delta^2}\int k^2\left(\frac{x}{\delta}\right)dx = \frac{1}{\delta}R(K)$$

Por tanto

$$C(K_\delta) = \left\{\delta^4 k_2^2 \frac{1}{\delta^4} R^4(K)\right\}^{1/5} = C(K).$$

El Kernel óptimo es aquel que minimiza

$$\int K^2(x)dx \quad s/t \qquad \sigma_K^2 = \sigma^2 = k_2$$

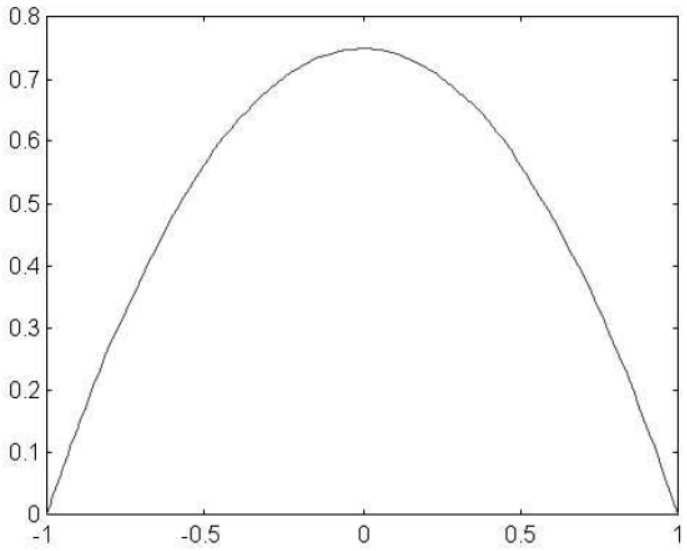
Verificándose que:

$$K \geq 0\,,\int K(x)dx = 1\,,\int xK(x)dx = 0\,,\int x^2K(x)dx = K_2 \neq 0$$

Hodges y Lehman (1956) demuestran que La función que minimiza La expresión anterior es La función de Epanechnikov

FIGURA 2

FUNCIÓN NÚCLEO DE EPANECHNIKOV



Parámetro de suavizado

El parámetro de suavizado controla el equilibrio que el estimador no paramétrico de la función de regresión debe mantener entre el buen ajuste a los datos observados y la capacidad de predecir bien observaciones futuras. Valores pequeños de h dan mucha flexibilidad al estimador y le permiten acercarse a todos los datos observador (cuando h tiende a 0 el estimador acaba por interpolar los datos), pero los errores de predicción asociados serán altos. Hay por lo tanto, sobre ajuste (*overfitting*). En el caso de que h tome un tamaño moderado no se ajustaría tan bien a las observaciones (tampoco es necesario, dado que los datos pueden contener ruido aleatorio) pero predeciría mejor. En otro extremo, si h es demasiado grande, tendremos falta de ajuste

(*underfitting*), como puede ocurrir con los modelo paramétricos globales.

Parámetro de suavizado en modelos no paramétricos

Los métodos de suavizamiento resultan muy útiles para la inferencia en los modelos de regresión. De hecho, la flexibilidad que caracteriza estos métodos permite descubrir las características que realmente subyacen en los datos disponibles. Sin embargo no siempre es fácil saber que características observadas en el suavizamiento están realmente ahí, o cuales son simplemente un artificio resultante de la variabilidad de los datos.

En este sentido una buena elección del parámetro de suavizado constituye un elemento crucial para obtener resultados en la práctica. La metodología clásica inferencial usando estimadores no paramétricos supone una elección del parámetro de suavizado, lo que habitualmente se hace compensando de forma optima las componentes de sesgo y de varianza del estimador. Esto, deriva de la minimización de algún criterio de error relacionado con el error cuadrático medio de la estimación resultante. En este caso el parámetro que determina el ancho de banda elegido constituiría una estimación de lo que se podría definir como parámetro de suavizado optimo y por tanto el objetivo de estimar a partir de los datos observados un buen

estimador para dicho parámetro optimo habitualmente desconocido.

Consideremos un problema de tipo de regresión no paramétrico multivariante

$$Y = m(X) + \varepsilon$$

Donde m se estimara por un suavizador d-dimensional. En general, este suavizador tiene la expresión:

$$\hat{m}(x: H) = n^{-1} \sum_{i=1}^{n} K_H(x - X_i) Y_i$$

Que depende de una función núcleo $, K, d - dimensional$, y de una matriz H, simétrica de dimensión $d * d$ denominada matriz de ancho de banda. La elección de la función núcleo resulta un problema de poca dificultad resuelto habitualmente utilizando elecciones que infieran buenas propiedades al estudio, mientras que la elección del parámetro de suavizado h, que conforma la matriz H, tiene una importancia crucial en el aspecto y propiedades del estimador de la función de regresión. En la práctica, valores distintos de h pueden producir estimadores completamente distintos.

Métodos de selección automática

Podemos clasificar los métodos de selección automática de parámetros de suavizado basados en una muestra en Métodos de primera generación y Métodos de segunda generación (JOMES, MARRON, & SHEATHER, 1996), Esta clasificación de

debe a la superioridad que han demostrado las técnicas desarrolladas a partir de 1990 (segunda generación), y las técnicas desarrolladlas con anterioridad a 1990 (primera generación). A continuación nos referimos a los métodos más utilizados y por ende estudiados.

- **Validación cruzada**

  Se caracteriza por utilizar la técnica de leave-one-out para minimizar alguna medida de discrepancia entre la densidad y su estimación.

  Validación cruzada de mínimos cuadrados LSCV (RUDEMO, 1982) (BOWMAN, An alternative method of cross validation for the smoothing of density estimates, 1984)

  Se basa en la minimización del MISE de la forma siguiente. Dado un estimador $\hat{f}$ de la densidad f, el MISE se expresa

$$MISE\hat{f} = E\int(\hat{f} - f)^2 = E\int \hat{f}^2 - 2E\int \hat{f}f + E\int f^2 \qquad (1.36)$$

  El ultimo termino no depende de la estimación $\hat{f}$, por tanto la elección de h para minimizar el $MISE$ equivale a la minimización de

$$\Phi(\hat{f}) = E\left[\int \hat{f}^2 - 2\int \hat{f}f\right] \qquad (1.37)$$

  Un estimador de $\int \hat{f}f(x)$ viene dado por

$$\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-1}(X_i) \tag{1.38}$$

Donde $\hat{f}_{-i}(x)$ es la densidad estimada a partir de los datos extrayendo de la muestra el dato X_i,

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x - X_j}{h}\right) \tag{1.39}$$

El estimador de (1.38) es un estimador insesgado para $E\int \hat{f}f$.Por lo tanto, un estimador insesgado de

$$\Phi(\hat{f}) = E\left[\int \hat{f}^2 - 2\int \hat{f}f\right] \tag{1.40}$$

Viene dado por

$$\hat{h}_{LSCV} = argmin_h LSCV(x) \tag{1.41}$$

Suponemos que el mínimo de (1.41) estara cercano al minimo de (1.40) y por lo tanto que el parámetro de ventana obtenido el minimizar (1.41) será una buena elección. Siendo (1.41) equivalente a

$$LSCV(h) = \frac{R(K)}{nh} + \frac{2}{n^2 h}\sum_{i<j}\gamma(c_{ij}) \tag{1.42}$$

Donde

$$\gamma(c) = (K * K)(c) - 2K(c) = \int K(w)K(w+c)dw - 2K(c) \tag{1.43}$$

Y $c_{ij} = \frac{X_i - X_j}{h}$ (1.44)

- **Tipo Plug- In**

Trata de substituir en la expresión del h_{AMISE} dada en (1.29) el valor $R(f'')$ a través de una muestra piloto. El problema es

escoger el parámetro de suavización para esta muestra piloto. Se han propuesto varias aproximaciones

Sheather y Jones (1991)

$$R(f'') = \int (f'')^2 dx = \int f'' df'(x) = -\int f'''(x) f'(x) dx$$

$$= -\int f'''(x) df(x) = \int f^{(4)}(x) f(x) dx = \psi_4 \quad (1.45)$$

Es decir

$$R(f'') = \psi_4 = E\left[f^{(4)}(x)\right]$$

Un posible estimador es

$$\psi_4(g) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}^{(4)}(X_i) = \frac{1}{n^2 g^5} \sum_i \sum_j K^{(4)}\left(\frac{X_j - X_i}{g}\right) \quad (1.46)$$

Definimos

$$\hat{h}_{DPI} = \left(\frac{R(K)}{k_2^2 \psi_4(g)}\right)^{1/5} n^{-1/5} \quad (1.47)$$

Evidentemente $\hat{h}_{DPI}$ depende de g (*pilot bandwith*) que puede ser obtenida de

$$g = \left(\frac{2K^{(4)}(0)}{k_2}\right)^{\frac{1}{7}} R(f''')^{-1/7} n^{-1/7} \quad (1.48)$$

Con $R(f''')$ estimado por un proceso análogo obteniendo una función ψ_6, y así sucesivamente hasta que finalmente se estima el termino $R(f^{(i)})$ tomando como referencia una distribución paramétrica como puede ser lo normal.

Generalmente no se realizan más de dos o tres procesos iterativos Jones y Sheather (1991).

Hall, Sheather, Jones y Marron (1991)

Trabajan con una mejor aproximación del $AMISE$, en particular mejoran la aproximación del sesgo, con lo que se obtiene

$$AMISE(h) = \frac{R(K)}{nh} - \frac{R(f)}{n} + \frac{1}{4}h^4 k_2^2 R(f'') - \frac{1}{24}h^6 k_2 k_4 R(f''') \quad (1.49)$$

Nótese que el segundo termino es constante, no depende de h y puede ser ignorado. Se demuestra que el minimizador de (1.49) viene dado por

$$\hat{h}_{PI} = \left(\frac{J_1}{n}\right)^{1/5} + \left(\frac{J_1}{n}\right)^{3/5} J_2 \quad (1.50)$$

Con

$J_1 = \frac{R(K)}{k_2^2 R_{h_1}(f'')}$ y $J_2 = \frac{k_4 R_{h_2}(f''')}{20 k_2 R_{h_1}(f'')}$ y estimando $R_{h_1}(f'')$ y $R_{h_2}(f''')$ por

$$\hat{R}_{h_1}(f'') = \frac{1}{n(n-1)h_1^5}\sum_{i,j} L^{(4)}\left\{\frac{X_i - X_j}{h_1}\right\}$$

Y $\hat{R}_{h_2}(f''') = \frac{-1}{n(n-1)h_1^7}\sum_{i,j}\phi^{(6)}\left\{\frac{X_i - X_j}{h_2}\right\}$

- **Métodos basados en aproximaciones bootstrap**

La idea básica es estudiar el $MISE$ a través de un versión bootstrap de la forma

$$MISE_*(h) = E_* \int \left(\hat{f}^*(t;h) - \hat{f}(t;g)\right)^2 dt \quad (1.51)$$

Donde E_* es la esperanza respecto a la muestra bootstrap $X_1^*, \ldots, X_n^*$, g es un parámetro de ventana piloto, $\hat{f}(t;g)$ una estimación de la densidad basada en la muestra original $X_1, \ldots., X_n$ y $\hat{f}^*(t;h)$ una estimación basada en la muestra bootstrap. Escogemos el valor $\hat{h}_{BT}$ que minimiza (1.51). Las diferencias entre las diferentes versiones radican en la elección de g y en la manera de generar la muestra bootstrap. Destaquemos que en Taylor (1989) utiliza $g = h$ llegando a la siguiente conclusión:

$$MISE_*(h) = \frac{1}{2n^2 h(2\pi)^{\frac{1}{2}}}\left(\sum_{i,j} exp\left\{-\frac{(X_i - X_j)^2}{8h^2}\right\} - \frac{4}{3^{\frac{1}{2}}}\sum_{i,j} exp\left\{-\frac{(X_i - X_j)^2}{6h^2}\right\}\right.$$

$$\left. +2^{1/2}\sum_{i,j} exp\left\{-\frac{(X_i - X_j)^2}{4h^2}\right\} + n2^{1/2}\right)$$

2.2.2 Análisis multivariante

El análisis multivalente es un conjunto de técnicas estadísticas cuya finalidad es analizar simultáneamente un conjunto de datos en el sentido de que hay varias variables para un individuo, una de las dificultades en definir que es el análisis multivariantes reside en el hecho de que el termino multivariantes (o multivariado) no ha sido usado de manera consistente en la literatura. Algunos investigadores usan el término multivariado simplemente para referirse a las relaciones existentes entre más de dos variables. Sin embargo, para que un análisis sea considerado verdaderamente multivariente, todas las variables

deben de ser aleatorias y deben de estar interrelacionadas de tal manera que los diferentes efectos no puedan ser interpretados significativamente de manera independiente (CAYUELA, 2010).

Normal multivariante

Consideramos en lo que sigue variables aleatorias $n-variables$, es decir, aplicaciones $X: \Omega \longrightarrow \mathbb{R}^n$. A cada $\omega \in \Omega$ corresponderá entonces un $X = X(\omega) \in \mathbb{R}^n$. Designaremos por $X_i = (X_{i1} X_{i2}, \dots, X_{in})'$ a la observación $i-ésima$ de la variable aleatoria $n-variante\ X$, y por $F_X(x) y\ f_X(x)$ a las funciones de distribución y densidad respectivamente de X. Emplearemos el convenio de utilizar mayúsculas para las variables aleatorias y minúsculas para sus valores concretos en un muestreo determinado. Llamaremos X_j a la variable aleatoria $j-ésima$.

Si estudiamos a estas variables por separado (con técnicas univariantes) perderíamos la oportunidad de estudiar la correlación entre diferentes variables $X_j\ y\ X_k$ en X (TUSELL, 2012). Los métodos de Análisis Multivariante comparten la idea de explorar esta información.

Llamaremos

$$u_X = EX \tag{2.1}$$

$$\sum_X = E\left[(X - u_X)(X - u_X)'\right] \tag{2.2}$$

Al igual que la distribución normal desempeña un papel destacado en la Estadística univariante, una generalización de ella, la distribución normal multivariante, constituye un modelo teórico de gran trascendencia en el Análisis Multivariante.

Distribución normal multivariante

Se dice que $X \sim N(0,1)$ si:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad -\infty < x < \infty$$

Y por ende:

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx \quad -\infty < x < \infty \tag{2.3}$$

$$\psi_X(u) = E e^{iuX} \tag{2.4}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-iu)^2} e^{-\frac{1}{2}u^2} dx \tag{2.5}$$

$$= e^{-\frac{1}{2}u^2} \tag{2.6}$$

Por transformación lineal de una variable aleatoria $N(0,1) = Y = \sigma X + u$ se obtiene una variable aleatoria normal general $N(u, \sigma^2)$ cuyas funciones de densidad, distribución y características son:

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-u)^2}{2\sigma^2}} \quad -\infty < y < \infty \tag{2.7}$$

$$F_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{(y-u)^2}{2\sigma^2}} dy \quad -\infty < y < \infty \tag{2.8}$$

$$\psi_Y(u) = e^{iu\mu - \frac{1}{2}\sigma^2 u^2} \tag{2.9}$$

Si tenemos p variables aleatorias X_j con distribución $N(0,1)$, independientes unas de otras, la función de densidad conjunta

de la variable aleatoria $p - variante\ X = (X_1, \ldots., X_p)'$ viene dada por el producto de las marginales

$$f_X(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}(x_l^2 + \cdots + x_p^2)} \tag{2.10}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}(x'Ix)} \tag{2.11}$$

Y la función característica por:

$$\psi_X(u) = e^{-\frac{1}{2}u'u} \tag{2.12}$$

Decimos que la variable aleatoria p-variable X cuya función de densidad es (2.10) sigue una distribución $N_p(\vec{0}, I)$ designado el primer argumento el vector de medias y el segundo la matriz de covarianzas. Esta última es diagonal, en virtud de la independencia entre las distintas componentes de X.

Si efectuamos una transformación lineal $X \to Y$ como

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p + \mu_1 \tag{2.13}$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p + \mu_2 \tag{2.14}$$

$$\vdots$$

$$Y_p = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p + \mu_p \tag{2.15}$$

o, en notación matricial, $Y = AX + \mu$, y A es de rango completo, tenemos que $X = A^{-1}(Y - \mu)$ y la función de densidad de Y se obtiene fácilmente de la de X:

$$f_Y(y) = f_X(A^{-1}(y - \mu))\left|\frac{\partial X}{\partial Y}\right| \tag{2.16}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}(y-\mu)'(A^{-1})'(A^{-1})(y-\mu)|A^{-1}|} \tag{2.17}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^{p} \frac{1}{|A|} e^{-\frac{1}{2}(y-\mu)'(AA')(y-\mu)} \tag{2.18}$$

Como

$$\sum\nolimits_{Y} = E(Y-\mu)(Y-\mu)' \tag{2.19}$$

$$= EAXX'A' \tag{2.20}$$

$$= AA' \tag{2.21}$$

Tenemos que la función de densidad (2.18) puede escribirse así:

$$f_Y(y) = \left(\frac{1}{\sqrt{2\pi}}\right)^{p} \frac{1}{|\Sigma_Y|^{1/2}} e^{-\frac{1}{2}(y-\mu)'\Sigma_Y^{-1}(y-\mu)} \tag{2.22}$$

ya que $|A| = \sqrt{|A||A|} = \sqrt{|A||A'|} = \sqrt{|\Sigma_Y|}$. Por otra parte, la función característica de Y es:

$$\psi_Y(u) = Ee^{iu'Y} \tag{2.23}$$

$$= Ee^{iu'(AX+\mu)} \tag{2.24}$$

$$= \psi X(A'u)e^{iu'\mu} \tag{2.25}$$

$$= e^{iu'\mu-\frac{1}{2}u'AA'u} \tag{2.26}$$

$$= e^{iu'\mu-\frac{1}{2}u'\Sigma_Y u} \tag{2.27}$$

La expresión (2.22) requiere para estar definida que Σ_Y sea de rango total sólo así puede encontrarse la inversa. La expresión (2.27) por el contrario es una función característica incluso aunque Σ_Y sea de rango diferente. Se dice que (2.22) y (2.27) son funciones de densidad y característica de un vector aleatorio con distribución $N_p(\mu, \Sigma_Y)$. Si $\Sigma_Y$es de rango deficiente, se dice que estamos ante una distribución normal singular, que carece de densidad (2.22)

Observación: La función de densidad normal multivariante es unimodal, alcanza su máximo para y coincidente con el valor de medias μ, y tiene contornos de igual densidad elípticos (o híper-elípticos).

Los siguientes hechos son de muy sencilla demostración:

Las distribuciones de cualesquiera combinaciones lineales de componentes de Y son normales.

Si Y es normal multivariante, cualesquiera marginales son normales uni-o-multivariantes.

Si $X\ e\ Y$ son vectores independientes conjuntamente definidos con distribuciones respectivas $N_p(\mu_X, \Sigma_X)$ y $N_p(\mu_Y, \Sigma_Y)$, y A, B son matrices cualesquiera de orden $d * p, (d \leq p)$, y rango d, se verifica:

$$AX + BY \sim N_d(A\mu_X + B\mu_Y, A\Sigma_X A' + B\Sigma_Y B')$$

Como caso particular, $CX \sim N_d(C\mu_X, C\Sigma_X C')$.

La incorrelación entre cualesquiera componentes X_i, X_j (o grupos de componente) de X, implica si independencia. En el caso de variables aleatorias con distribución normal multivariante, incorrelación e independencia son nociones coextensivas.

Transformaciones lineales ortogonales de vectores $N_d(\vec{0}, \sigma^2 I)$ tiene distribución $N_d(\vec{0}, \sigma^2 I)$.

Observación: Una normal multivariante tiene contornos de igual densidad, cuando esta densidad existe, cuya expresión viene dada por:

$$-\frac{1}{2}(y-\mu)'\Sigma_Y^{-1}(y-\mu)=k$$

Como la matriz de covarianzas (en el caso de rango completo, para el que existe la densidad) es definida positiva, la expresión anterior proporciona la superficie de un híper-elipsoide: un elipse ordinaria en $\mathbb{R}^2$, un elipsoide (similar al un balón de rugby) en $\mathbb{R}^2$, y figuras que ya no podemos visualizar en más de tres dimensiones.

Observación: Hay versiones multivariantes del Teorema del Limite Central, que sugieren que variables multivariantes que son:

Suma de muchas otras.

Aproximadamente independiente Y.

Sin influencia abrumadora de ninguna sobre el conjunto, siguen una distribución aproximadamente normal multivariante. Es un hecho, sin embargo, que el supuesto de normalidad multivariante es sumamente restrictivo (TUSELL, 2012), y de dar plausibilidad en la práctica. En particular, el supuesto de normalidad multivariante es mucho más fuerte que el de normalidad de las marginales.

Modelos de probabilidad No Lineal

La estimación e interpretación de los modelos probabilísticos lineales plantea una serie de problemas que ha llevado a la búsqueda de otros modelos alternativos que permitan estimaciones más fiables de las variables dicotómicas. Para evitar la variable endógena estimada puede encontrarse fuera del rango $(0, 1)$ (MEDINA, 2003), las alternativas disponibles son utilizar modelos de probabilidad no lineales, donde la función de especificación utilizada garantice un resultado en la estimación comprendido en el rango $0 - 1$. Las funciones de distribución cumplen este requisito, ya que son funciones continuas que toman valores comprendidos entre 0 y 1.

Dado que el uso de una función de distribución garantiza que el resultado de la estimación este acotado entre 0 y 1 , en principio son varias las alternativas, siendo las más habituales la función de distribución logística, que ha dado lugar al modelo Logit, y la función de distribución de la normal tipificada, que ha dado lugar al modelo Probit. Tanto los modelos Logit como Probit relacionan, por tanto, la variable endógena Y_i con las variables explicativas, x_{ki} a través de una función de distribución.

En el caso del modelo Logit, la función utilizada es la logística, por lo que la especificación de este tipo de modelos es la siguiente:

$$Y_i = \frac{1}{1 + e^{-a - B_i x_{ki}}} + \varepsilon_i = \frac{e^{a + B_i X_{ki}}}{1 + e^{a + B_i X_{ki}}} + \varepsilon_i$$

En el caso de modelo Probit la función de distribución utilizada es la de la normal tipificada, con lo que el modelo queda especificado a través de la siguiente expresión:

$$Y_i \int_{-\infty}^{a + B x_i} \frac{1}{(2\pi)^{1/2}} e^{-\frac{s^2}{2}} ds + \varepsilon_i$$

Donde la variable s es una variable "muda" de integración con media cero y varianza uno.

Dada la similitud existente entre las curvas de la normal tipificada y de la logística, los resultados estimados por ambos modelos no difieren mucho entre sí (MEDINA, 2003), siendo las diferencias operativas, debidas a la complejidad que presenta el cálculo de la función de distribución normal frente a la logística, ya que la primera solo puede calcularse en forma de integral. La menor complejidad de manejo que caracteriza el modelo Logit es lo que ha potenciado su aplicación en la mayoría de los estudios empíricos.

Regresión logística binaria

Esta técnica se originó en la década de los 60 (DOMINGUEZ & ALDANA, 2001). El objeto de esta técnica estadística es expresar la probabilidad de que ocurra un hecho como función de ciertas variables, a través de una ecuación lineal que consiga explicar la máxima variación posible de "y".

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_i x_i$$

Supongamos que son k-variables $(k > 1)$, que se consideran potencialmente influyentes en la variable y. La regresión logística como cualquier otra técnica de regresión ofrece la posibilidad de evaluar la influencia de cada variable independiente sobre la variable dependiente (respuesta, la cual puede ser dicotómica o politómica según sea el caso) y controlar el resto de variables.

Al ser la variable dicotómica, podrá solamente tomar dos valores "1", en caso de que el hecho (presencia del Síndrome de Burnout ocurra), y "0", en el caso contrario.

Se construye un modelo que permita prever el valor de la variable ficticia binaria de un elemento de una población, en función de ciertas características medibles x. Supongamos que se dispone de una muestra de n elementos del tipo (y_i, x_i), donde y_i es igual a 0 cuando el elemento pertenece a la primera población P_1 y 1 cuando pertenece a la segunda P_2. A su vez, x_i es un vector de variables explicativas.

El primer enfoque es formular el siguiente modelo de regresión:

$$y = \beta_0 + \beta_1' x + u$$

Y estimar los parámetros por mínimos cuadrados de la forma habitual. Este método es equivalente a la función lineal discriminante de Fisher. Este procedimiento es óptimo para

clasificar si la distribución conjunta de las variables explicativas es normal multivariante, con la misma matriz de covarianzas. Sin embargo, la discriminación lineal puede funcionar mal en otros contextos, cuando las covarianzas sean distintas o las distribuciones muy alejadas de la normal. Además, si un objetivo importante del estudio es identificar que variables son mejores para clasificar entre las dos poblaciones, la función lineal se encuentra con problemas de interpretación, tanto del modelo como de los coeficientes estimados.

En concreto, tomando esperanzas para $x = x_i$

$$E[y|\mathrm{x_i}] = \beta_0 + \beta_1' x_i$$

Llamamos p_i a la probabilidad de que y tome el valor 1 cuando $x = x_i$

$$p_i = P(y = 1|\mathrm{x_i})$$

Y la esperanza de y es:

$$E[y|\mathrm{x_i}] = P(y = 1|\mathrm{x_i}) * 1 + P(y = 0|\mathrm{x_i}) * 0 = \mathrm{p_i}$$

por tanto,

$$p_i = \beta_0 + \beta_1' x_i$$

Que es una expresión equivalente del modelo. En consecuencia, la predicción $\hat{y}_i$ estima la probabilidad de que un individuo con

características definidas por $x = x_i$ pertenezca a la población correspondiente a $y = 1$.

El inconveniente principal de esta formulación es que p_i debe estar entre cero y uno, y no hay ninguna garantía de que la predicción $\beta_0 + \beta_1' x_i$ verifique esta restricción, ya que el modelo puede prever probabilidades mayores que la unidad. Esto no es un problema insalvable para clasificar, pero lo es si queremos interpretar el resultado de la regla de clasificación como una probabilidad de pertenencia a cada población.

A pesar de este inconveniente, este modelo simple conduce a una buena regla de clasificación, ya que según la interpretación de Fisher, maximiza la separación entre los grupos, sea cual sea la distribución de datos. Sin embargo, cuando los datos no son normales, o no tienen la misma matriz de covarianzas, la clasificación mediante una ecuación de relación lineal no es necesariamente óptima, el modelo logístico puede conducir a mejores resultados (SALCEDO).

El modelo Logit

Si queremos que el modelo proporcione directamente la probabilidad de pertenecer a cada uno de los grupos, debemos transformar la variable respuesta de algún modo para garantizar que la respuesta prevista este entre cero y uno. Si tomamos,

$$p_i = F(\beta_0 + \beta_1' x_i)$$

Garantizaremos que p_i esté entre cero y uno si exigimos de $F(.)$ tenga esta propiedad.

La clase de funciones no decrecientes, acotadas entre cero y uno, es la clase de las funciones de distribución, por lo que el problema se resuelve tomando como $F(.)$ cualquier función de distribución.

Habitualmente se toma como $F(.)$ la función de distribución logística, dada por:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1' x_i)}}$$

Esta función tiene la ventaja de ser continua. Además, como,

$$1 - p_i = \frac{e^{-(\beta_0 + \beta_1' x_i)}}{1 + e^{-(\beta_0 + \beta_1' x_i)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1' x_i)}}$$

Resulta que

$$g_i = \log \frac{p_i}{1 - p_i} = \log \left(\frac{\frac{1}{1+e^{-(\beta_0 + \beta_1' x_i)}}}{\frac{e^{-(\beta_0 + \beta_1' x_i)}}{1+e^{-(\beta_0 + \beta_1' x_i)}}} \right) = \log \left(\frac{1}{1+e^{-(\beta_0 + \beta_1' x_i)}} \right) = \beta_0 + \beta_1' x_i$$

(2.28)

De modo que, al hacer la transformación, se tiene un modelo lineal que se denomina ***logit,*** lo importante de esta transformación es que tiene muchas propiedades semejantes al Modelo de Regresión Lineal Simple, por ejemplo sus parámetros, puede ser continua y puede tomar cualquier valor real dependiendo de x (SALCEDO).

La variable g representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a ambas poblaciones y, al ser una función lineal de las variables explicativas, facilita la estimación y la interpretación del modelo.

Una ventaja adicional del modelo **logit** es que si las variables son normales verifican el modelo **logit**, y además, también es cierto que una amplia gama de situaciones distintas a la normal.

En efecto, si las variables son normales multivariantes

$$g_i = \log\frac{f_1(x)}{f_2(x)} = -\frac{1}{2}(x-u_1)'V^{-1}(x-u_1) + \frac{1}{2}(x-u_2)'V^{-1}(x-u_2)$$

Simplificando,

$$g_i = \frac{1}{2}(u_2V^{-1}u_2 - u_1V^{-1}u_1) + (u_1-u_2)'V^{-1}x$$

Por lo tanto, g_i es una función lineal de las variables x. Comparando con (2.28) la ordenada en el origen β_0 es igual a

$$\beta_0 = -\frac{1}{2}\omega'(u_1+u_2)$$

Donde $\omega = V^{-1}(u_1-u_2)$, y el vector de pendientes es $\beta_1 = \omega$.

Función de verosimilitud

Con el fin de estimar β y analizar el comportamiento del modelo considerado, observamos una muestra aleatoria simple de tamaño n dada por $\{(x_i', y_i): i = 1, \dots, n\}$ donde $x_i =$

$(x_{i1}, \dots, x_{ik})$'es el valor de las variables independiente e $y_i = \{0,1\}$ es el valor observado de Y en el i-ésimo elemento de la muestra.

$$Y/(x_1, \dots, x_k) \sim Binomial(1, p\left(Y = \frac{1}{X_1, .., X_k}; B\right))$$

Utilizando el hecho de que la variable dependiente toma sólo dos resultados (éxito ó fracaso; 0 ó 1), cuando el número de éxitos en n repeticiones tiene una distribución binomial $B(n, p)$.

La función de verosimilitud es:

$$L(B/(X_1, Y_1), \dots (X_n, Y_n)) = \prod_{i=1}^{n} p_i{}^{y1}(1 - pi)^{1-y1}$$

Donde: $p_i = p(x_i' : B) = p(x_{i1}, \dots, x_{ik}; B); i = 1, \dots, n$

Estimación de los coeficientes de regresión logística

El cálculo de estos coeficientes es complejo, esta estimación se puede realizar mediante diversos métodos, pero el más utilizado es el de máxima verosimilitud (CACERES, 2007).

Esta función es la siguiente:

$$Lo(B) = P(Y)^{\sum_{j=1}^{n} y_j} * (1 - P(Y))^{(n - \sum_{j=1}^{n} y_j)}$$

También se suele utilizar el logaritmo neperiano

$$LLo(B) = \sum_{j=1}^{n} y_j * \ln\big(P(Y)\big) + \left(n - \sum_{j=1}^{n} y_j\right) * \ln(1 - P(Y))$$

Es la expresión anterior $y_j$es el j-esimo valor de la variable Y, que es 1 si ocurre el suceso de interés y 0 si no ocurre, B es el coeficiente de regresión logística si es un modelo simple, si es

un modelo múltiple es el vector de coeficientes de regresión logística: $(B_1, B_2, \dots, B_k)$.

Un parámetro muy utilizado en regresión logística es $-2.LLo(B)$ se utiliza como estadístico de contraste de hipótesis de modelo de regresión logística.

Los valores de los estimadores de regresión logística son aquellas que hacen máxima la función de verosimilitud o su logaritmo, el cálculo se hace mediante el método de Newton-Raphson, que es un método de derivaciones sucesivas e interactivas. Los cálculos son complejos, sobre todo en el modelo múltiple en la que algunos de los elementos de las ecuaciones son matrices. No todos los conjuntos de datos pueden ajustarse a un modelo de regresión logística, los programas informáticos hacen un número determinado de interacciones, y si no se consigue la convergencia, se da un mensaje de no convergencia de los datos con el modelo logístico.

Una vez calculados los coeficientes de regresión logística con la ayuda de programas informáticos, el cálculo de las probabilidades es más sencillo, los datos deben proceder de una muestra aleatoria, y el cálculo de las probabilidades son estimaciones a las probabilidades poblacionales.

Regresión local logística

Una característica importante de la función de regresión es que se trata de una función de probabilidad condicional, y por lo tanto únicamente toma valores entre 0 y 1. Esta propiedad la verifica también el estimador Nadaraya- Watson $p_n^{NW}(.)$, pero no el estimador local lineal $p_n^{LL}(.)$ que puede facilmente tomar valores fuera de ese intervalo.

Un ajuste de la función $p(.)$ que mantenga la flexibilidad del estimador local lineal, es decir, que no estime $p(.)$ localmente como una constante tal como hace el estimador de Nadaraya-Watson, pero que siempre tome valores entre $0\ y\ 1$ puede ser el estimador local logístico. Para estimar la función $p(.)$ en un punto $t \geq 0$, se ajusta localmente una función logística en un entorno de t, cuya amplitud dependerá de la ventana b.

La función de regresión logística es el ajuste más común para datos binarios. Consiste en ajustar la función $logit$ de la probabilidad de la variable respuesta como una función lineal de la forma:

$$logit[p(t)] = log\left[\frac{p(t)}{1-p(t)}\right] = a + bt$$

Es decir,

$$p(t) = \frac{\exp(a+bt)}{1-\exp(a+bt)}$$

El ajuste logístico local consiste entonces en minimizar la siguiente función objetivo:

$$\sum_{i=1}^{n} l(Z_i, \delta_i, g) K\left(\frac{Z_i - t}{b}\right) \tag{2.28}$$

Donde $g(t) = exp(a - bt)(1 + \exp(a + bt)]^{-1}$, y el funciona $l(.,.,.)$puede ser

$$l(z, d, g) = -log[g(z)^d (1 - g(z))^{1-d}]$$

Que da lugar al estimador local logístico de máxima-verosimilitud $p_n^{LLOg\_MV}(.)$, o

$$l(z, d, g) = (d - g(z))^2$$

Que corresponde con el ajuste mínimo cuadrático $p_n^{LLog_{MC}}(.)$.

El método estándar para resolver de forma global estas ecuaciones es el procedimiento *local scoring*, es decir, el algoritmo de Newton- Raphson usando la matriz de información esperada en vez de la observada. Sin embargo, este algoritmo no se puede usar en el ajuste local. En este caso, es necesario calcular para cada punto $t > 0$ de estimación la funcion objetico dad en (2.28) y calcular los parámetros $a\ y\ b$ que la minimizan. Puesto que las ecuaciones no son lineales en los parámetros, las soluciones locales tienen que ser aproximadas de forma iterativa.

2.2.3 Análisis de supervivencia de datos

Al trabajar con datos dependientes del tiempo, los cuales están más presentes en salud, muchas veces se pierde la información

dado que muchos de los pacientes mueren, se trasladan de lugar para recibir otros tratamientos, o sus cuerpos rechazan los tratamientos proporcionados.

Este tipo de análisis lo que verifica es el estudio de los sujetos hasta la aparición de un suceso de interés y de las variables que pueden influir en la aparición del mismo, por lo que nos puede interesar saber si la presencia del suceso de interés se debe a factores como la edad de los individuos (BOTELLA, ROCAMORA, MARTÍNEZ, & ALACRÉU, 2001). El principio fundamental del porque no se modeliza a las variables desde una perspectiva de Regresión Lineal, es porque básicamente el tiempo de estudio no se comporta como una variable normal (CARREÑO, 2006), y además en ciertas ocasiones no se observa en los individuos el suceso de interés en el tiempo programado para el seguimiento de las muestras.

En ciertos casos se logra observar el suceso de interés en los individuos dentro del periodo de seguimiento de las muestras, en otros casos solo se tiene información parcial acerca del individuo, lo cual puede deberse a factores externos al estudio, como por ejemplo la muerte del individuo, la falta de tiempo para poder realizar el control y la toma de información, etc. Justamente a esta clase de datos de los cuales solo se tiene información parcial reciben el nombre de datos censurados.

Pueden ser diferentes las causas de censura pero siempre responden a alguno de los siguientes criterios:

Existe una pérdida del seguimiento (por ejemplo el individuo desaparece o decide no seguir colaborando con el estudio).

El estudio termina antes de que se produzca el evento.

Se produce otro evento que impide que ocurra el evento de interés (por ejemplo se observa la evolución de una enfermedad en un paciente desde su diagnostico hasta la muerte del individuo, pero el paciente fallece en un accidente de tránsito).

Censura de datos

El término de censura hace referencia a un tipo de pérdida de información en situaciones en las que la variable de interés es un tiempo de vida. La censura surge en las ocasiones en las que hay individuos de la muestra para los que no se conoce exactamente su tiempo de vida, sino que únicamente se sabe que éste ha ocurrido dentro de un cierto intervalo de valores. De esta forma se pueden considerar varios tipos de censura: censura por la derecha, por la izquierda y censura dentro de un intervalo. Si la variable de interés es el tiempo transcurrido desde que ocurre un suceso inicial hasta que ocurre un suceso final o suceso de interés, comúnmente llamado fallo, entonces se asume que hay censura por la derecha; cuando en el momento en que finaliza el estudio hay sujetos para los que no se conoce el instante exacto de fallo, sino que solamente se conoce que ha

sido posterior a un momento dado. Por tanto, el valor exacto del tiempo de vida será superior al valor observado (JÁCOME, 2005). Este problema es habitual cuando se analizan tiempos de vida, puesto que los estudios pueden terminar antes del fallo de todos los individuos de la muestra. Aquellos sujetos que no hayan fallado antes del fin del estudio serán censurados por la derecha, puesto que sólo sabremos que, de continuar el estudio, su instante de fallo sería posterior al instante final del estudio, y por tanto su tiempo de vida sería mayor que el observado. Lo mismo ocurre cuando no se puede observar el instante de fallo debido a la pérdida de seguimiento del individuo. Esto puede ser, entre otras causas, por un fallo debido a alguna razón ajena a la de interés, el abandono del estudio por parte del individuo, cambio de domicilio, etc.

Análogamente, el tiempo de vida asociado a un individuo en estudio se considera censurado por la izquierda si es menor que cierto valor dado, es decir, si el momento exacto en el que ocurrió el fallo es desconocido, sabiendo tan sólo que ha ocurrido antes de que el individuo se incluya en el estudio. Por ejemplo, es posible encontrarse en la muestra con sujetos que ya hayan fallado antes del comienzo del estudio, sin saber exactamente cuándo. Un tipo más general de censura que generaliza los dos anteriores surge cuando de alguno de los tiempos de vida sólo se conoce que pertenecen a cierto intervalo. Se habla entonces de censura de tipo intervalo. Es

común en estudios donde se hace un seguimiento periódico a los individuos. En este caso, cuando se encuentra un fallo para un individuo, solamente se sabe que el suceso de interés, el fallo, ocurrió entre dos revisiones periódicas.

De todos estos tipos de censura, el más común y en el que nos centraremos a partir de ahora, es el de censura por la derecha (JÁCOME, 2005). Ahora bien, como hemos visto en los ejemplos anteriores, las causas que originan la censura de una observación pueden ser aleatorias o controladas; esto hace que se distinga entre tres clases de censura:

Censura tipo I: El suceso se observa si ocurre antes de un momento fijo predeterminado C. En este caso, C es una constante prefijada por el investigador para todas las unidades muéstrales. Este tipo de censura es común cuando, por diversas causas, el investigador finaliza el estudio antes de que todos los individuos hayan experimentado el suceso de interés. Si no hay pérdidas accidentales, todas las observaciones censuradas son iguales a la longitud del periodo en estudio o tiempo calendario.

Censura tipo II: Este tipo de censura surge cuando se fija el final del estudio en el momento en que un número $r < n$ predeterminado de individuos falla. Los tiempos de vida observados son los r menores valores de la muestra, de forma que C se convierte en la variable aleatoria $C = T(r)$.

Censura tipo III: En la mayoría de los estudios, se fija la duración y los individuos entran a formar parte de la muestra a lo largo de ese periodo. Para los individuos que fallan antes del final del estudio, se conocen exactamente sus tiempos de vida. Para los que no han experimentado el suceso al final del estudio, la censura de sus tiempos de vida es semejante a la de tipo I. En ocasiones, algunos sujetos experimentan otros sucesos independientes del de interés que provocan su eliminación del estudio. Esta situación se denomina también censura aleatoria. En este tipo de censura, C es una variable aleatoria que se supone independiente de la variable de interés.

Modelo de censura aleatoria por la derecha

Se denota el tiempo de vida por Y con una función de densidad $f(.)$ y una función de distribución $F(.)$, lo que realmente se observa cuando se tiene la presencia de la cesura aleatoria por la derecha es una variable bidimensional (Z, δ), definida por:

$$Z = \min(Y, C) \quad \text{Y} \quad \delta = 1_{\{Y \leq C\}}$$

Siendo C la variable de censura por la derecha y 1_A la función indicadora del suceso A, es decir

$$z = \begin{cases} Y \; si \; Y \leq C \\ C \; si \; Y \geq C \end{cases} \quad \text{Y}$$

$$\delta = \begin{cases} 1 \; si \; la \; observacion \; no \; es \; censurada \; (Z = Y) \\ 0 \; si \; la \; observacion \; es \; censurada \; (Z = C) \end{cases}$$

Bajo este modelo C es una variable aleatoria con función de distribución $G(t) = P(C \leq t)$y la función de densidad $g(.)$, que supondremos independiente de Y. Deduciendo podemos afirmar que Z tiene una función de distribución $H(.)$ de la siguiente manera:

$$1 - H(t) = \big(1 - F(t)\big)(1 - G(t)) \qquad (3.1)$$

Supondremos además que las variables aleatorias son positivas, y adoptaremos la siguiente notación:

$a_F = inf\{t > 0 : F(t) > 0\}$ y $d_F = \sup\{t > 0 : F(t) < 1\}$.

Para representar los extremos inferior y superior del soporte de la función de distribución $F(.)$. Extenderemos esta notación no sólo para la función $F(.)$, sino también para las funciones de distribución $G(.)$ y $H(.)$: a_G, b_G, a_H, b_H respectivamente. Bajo la relación (3.1), estos extremos verifican lo siguiente:

$$a_H = mín\{a_F\,, a_G\}$$

$$b_H = mín\{b_F\,, b_G\}$$

Asociadas a este modelo necesitaremos definir algunas funciones de interés. En primer lugar, sea $p(t)$ la probabilidad condicional de que una observación sea no censurada a que $Z = t$, $p(t) = P(\delta = 1 \mid \mathrm{Z} = \mathrm{t}) = \mathbb{E}(\delta \mid \mathrm{Z} = \mathrm{t})$.

Junto a esta probabilidad, definimos ahora la probabilidad incondicional γ de que una observación sea censurada:

$\gamma = P(\delta = 1) = \mathbb{E}(\delta) = P(Z \leq C) = \int_{a_F}^{\infty}(1 - G(v))dF(v) = H^1(+\infty)$.

Donde $H^1(t) = P(Z \leq t, \delta = 1)$ es la función de subdistribución de las observaciones sin censura, que se puede escribir de la siguiente forma:

$$\int_{a_H}^{t} p(\upsilon)dH(\upsilon) = P(Z \leq t, \delta = 1) = H^1(t) = P(Y \leq t, Y \leq C) =$$

$$\int_{a_F}^{t}\left(1 - G(\upsilon^-)\right)dF(\upsilon),$$

Es decir, $dH^1(t) = p(t)dH(t)$.

Análogamente, la función de subdistribución de las observaciones censuradas es:

$$H^0(t) = P(Z \leq t, \delta = 0).$$

Estimación de la función de distribución

El estimador Kaplan Meier

Es un estimador paramétrito de la función de supervivencia para datos censurados aleatoriamente por la derecha más usado y estudiado en la práctica. También conocido en la literatura como estimador limite- producto (KAPLAN & MEIER, 1958). Su expresión se puede derivar siguiendo distintos métodos, algunos más intuitivos que otros, por lo que en la presente investigación tomaremos como referencia la obtenida a través de la relación entre la función de supervivencia y la razón de fallo acumulada:

$$1 - F(t) = exp[-\Lambda_F(t)] \qquad (3.2)$$

Esta relación se puede generalizar, al caso en el que la función $\Lambda_F(.)$ presente discontinuidades, de la siguiente manera:

$$1 - F(t) = \exp(-\Lambda_F^a(t)) \prod_{\{a_i \in A / a_i \leq t\}} (1 - \Lambda_F\{a_i\})$$

Donde $-\Lambda_F^a(.)$ denota la parte continua de $\Lambda_F(.)$, A el conjunto de puntos donde $\Lambda_F(.)$ tiene discontinuidades de salto y $\Lambda_F\{a_i\} = \Lambda_F\{a_i\} - \Lambda_F\{a_i^-\}$ es la magnitud del salto de $\Lambda_F(.)$ en a_i (SHORACK & WELLNER, 1986).

En este modelo de censura podemos escribir la razón de fallo acumulada $\Lambda_F(.)$ en función de cantidades estimables empíricamente de la siguiente manera:

$$\Lambda_F(t) = \int_0^t \lambda(v)dv = \int_0^t \frac{dF(v)}{1-F(v^-)} = \int_0^t \frac{1-G(v^-)}{1-H(v^-)} dF(v) = \int_0^t \frac{dH^1(v)}{1-H(v^-)} =$$

$$\int_0^t \frac{p(v)}{1-H(v^-)} dH(v)$$

Para todo $t < b_H$. Si sustituimos las funciones $H(.)$ y $H^1(.)$ en la expresión anterior por sus respectivas estimaciones empíricas, $H_n(t) = \frac{1}{n}\sum_{i=1}^n Z_{\{Z_i \leq t\}}$ y $H_n^1(t) = \frac{1}{n}\sum_{i=1}^n 1_{\{Z_i \leq t\}}\delta_i$.

Y tomamos $H_n(t^-) = \lim_{x \to t} H_n(x)$, obtenemos el estimador no parametrico de la razón de fallo acumulada más utilizando en este contexto, el conocido estimador de Nelson Aalen:

$$\Lambda_n^{NA}(t) = \int_0^t \frac{dH_n^1(v)}{1-H_n(v^-)} = \sum_{i=1}^n \frac{1_{\{Z_i \leq t, \delta_i = 1\}}}{n(1-H_n(Z_i^-))} = \sum_{Z_{(i)} \leq t} \frac{\delta_{[i]}}{n-i+1} \qquad (3.3)$$

Donde $\{Z_{(i)}, \delta_{[i]}\}_{i=1}^{n}$ son las observaciones ordenadas, y $\delta_{[i]}$'s son los concomitantes correspondientes a los individuos de no censura. Notemos que precisamente $n\delta_{[i]}/(n-i+1)$ es la estimación empírica de $p(v)/(1-H(v^{-}))$.

Este estimador fue sugerido por primera vez por Nelson (1972) en el contexto de la fiabilidad, y redescubierto por Aalen (1978), quien lo obtuvo usando técnicas de procesos de contar (ADERSEN, BORGAN, GILL, & KEIDING, 1993) (FLEMING & HARRINGTON, 1991).

Se puede modificar la expresión (3.3) del estimador de Nelson-Aalen para permitir más de un fallo en un instante t. Supongamos que los sucesos ocurren en D tiempos distintos $t_1 < \cdots < t_p$, y que en el instante t_i las d_i sucesos o fallos, siendo N_i el número de individuos en riesgo, es decir, el número de individuos vivos en t_i. En el caso de que solo haya un fallo en cada instante, entonces $d_i = \delta_{[i]}$ y $N_i = n - i + 1$.

El cociente d_i/N_i proporciona una estimación de la probabilidad condicionada de que un individuo que sobrevive hasta justo antes del instante t_i, falle en el instante t_i. Esta es la cantidad básica a partir de la cual se construyen los estimadores de la función de supervivencia y de la razón de fallo acumulada. En este caso, el estimador de Nelson – Aalen se puede escribir de la forma:

$$\Lambda_n^{NA(2)}(t) = \sum_{Z_{(i)} \le t} \frac{d_i}{N_i}.$$

A partir de la igualdad (3.2), resulta el siguiente estimador de la función de supervivencia en el instante t:

$$1 - \hat{F}(t) = exp[-\Lambda_n^{NA}(t)] = exp\left[-\sum_{z_{(i)} \le t} \frac{\delta_{[i]}}{n-i+1}\right] = \prod_{Z_{(i)} \le t} exp\left(-\frac{\delta_{[i]}}{n-i+1}\right)$$

Podemos obtener una expresión de este estimador más cómoda de calcular si usamos la aproximación $e^{-t} \simeq 1 - t$ para t próximos a 0, dando lugar a la expresión final del estimador

Kaplan y Meier de la función de supervivencia:

$$1 - F_n^{KM}(t) = \prod_{Z_{(i)} \le t} \left(1 - \frac{\delta_{[i]}}{n-i+1}\right)$$

Análogamente, en el caso de permitir más de un fallo en un instante t, el estimador de la función de supervivencia sería:

$$1 - F_n^{KM(2)}(t) = \prod_{Z_{(i)} \le t} \left(1 - \frac{d_i}{N_i}\right)$$

Se puede comprobar que el estimador de Nelson – Aalen de la razón de fallo acumulada es el primer término de la serie de Taylor menos el logaritmo del estimador de Kaplan Meier de la función de supervivencia. En concreto, la relación entre los estimadores de Kaplan- Meier y Nelson Aalen es:

$$1 - F_n^{KM}(t) = exp[-\Lambda_n^{NA}(t)] + O_p(n^{-1})$$

Propiedades del estimador Kaplan Meier

Este estimador presenta propiedades destacables, su facilidad de cálculo y el hecho de que sea un estimador no paramétrito de máxima verosimilitud para datos censurados (JOHANSEN, 1978), estas propiedades hacen que sea el estimador más utilizado y estudiado en este contexto. Sin embargo presenta ciertos problemas cuando la hipótesis de independencia entre los tiempos de fallo Y y los tiempos de censura C no se verifica, por otro lado tiene saltos escalonados únicamente en las observaciones no censuradas, con pesos que aumenten desde el menor al mayor dato no censurado, puesto que dependen del número de observaciones censuradas entre ellos (EFRON, 1967).

El principal resultado para la versión de Efron del estimador límite producto es el siguiente:

- **Propiedad 1** (Teorema 1 de Phadia y Shao, (PHADIA & SHAO, 1999))

  Sea $\bar{F}_n^{KM}(.)$ el estimador de Kaplan- Meier de la función de supervivencia $\bar{F}(.) = 1 - F(.)$. Entonces, el momento k-esimo de $\bar{F}_n^{KM}(.)$ viene dato por

  $$\mathbb{E}[(\bar{F}_n^{KM})^k(t)] = \sum_{i=0}^{n-1} \frac{n!}{(n-i)!} \bar{H}^{n-i}(t) \int_0^t \int_0^{ti} \dots \int_0^{t2} \prod_{j \le i} d\varphi_j(t_j) \quad (3.4)$$

  Donde $0 < t_1 < t_2 < \cdots < t_i \le t$ y

  $$\varphi_j(t) = H(t) - H^1(t) + H^1(t) \left(\frac{n-j}{n-j+1}\right)^k ,$$

Siendo $H^1(t) = p(Z \leq t, \delta = 1)$. El producto se calcula sobre $j = 1,2,\ldots.,i$ para $i = 0$, el producto se define como 1.

- **Propiedad 2** (Teorema 2 de Phadia y Shao, (PHADIA & SHAO, 1999))

  Si se aproxima cada función $\varphi_j(x)$ por su componente lineal $\left[\left(\varphi_j(t) - \varphi_j(0)\right)/t\right]x$, en el intervalo $(0,t)$, y se sustituye en la expresión (3.4), se obtiene

$$\mathbb{E}[(\bar{F}_n^{KM})^k(t)] \simeq \sum_{i=0}^{n-1} \binom{n}{i} \bar{H}^{n-i}(t) \prod_{j \leq i} \left[H(t) - H^1(t) + H^i(t)\left(\frac{n-j}{n-j+1}\right)^k\right]$$

- **Propiedad 3** (Normalidad asintótica y puntual y sobre intervalos compactos, (FÖLDES & RETJÖ, 1980))

  Sean las funciones de distribución $F(.)$ y $G(.)$ continuas. Entonces:

  Para todo $0 < t < b_H$,

$$\sqrt{n}\left(F_n^{KM}(t) - F(t)\right) \underrightarrow{d} N(0, \sigma(t))$$

  Donde

$$\sigma^2(t) = (1 - F(t))^2 \int_0^t \left(1 - H(\upsilon)\right)^{-2} dH^1(\upsilon)$$

  El proceso estocástico $\sqrt{n}(F_n^{KM} - F)$ converge globalmente de $D[0.T]$ para cada $T < b_H$ a un proceso Gaussiano $Z(.)$

$$\sqrt{n}(F_n^{KM} - F) \xrightarrow{d} Z(.)$$

  Con media 0 y función de covarianzas:

$$\mathrm{Cov}\left(\mathrm{Z(x)}, \mathrm{Z(t)}\right) = \left(1 - \mathrm{F(x)}\right)(1 - \mathrm{F(t)}) \int_0^{\mathrm{x} \wedge \mathrm{t}} \left(1 - \mathrm{H}(\upsilon)\right)^{-2} \mathrm{dH}^1(\upsilon),$$

  Siendo $D[0,T] = \{f \in F([0,T], \mathbb{R})$: f *continua por la derecha y*

$discontinuidades, a\ lo\ sumo, de\ salto\}$, con la topología de Skorohod, y $F([0,T],\mathbb{R})$ el conjunto de las funciones que van de $[0,T]$ a $\mathbb{R}$.

- **Propiedad 4** (Consistencia uniforme fuerte, (CSÖRGO & HORVÁTH, 1983))

  Sea $0 < T < b_H$. Entonces

  $$|F_n^{KM} - F| \longrightarrow 0 \text{ c. s. uniformemente en } [0,T]$$

  Si además $G(b_F^-) > 0$, donde $b_F = sup\{t: F(t) < 1\}$, entonces:

  $|F_n^{KM} - F| \longrightarrow 0$ c.s. uniformemente en $\mathbb{R}$

- **Propiedad 5** (Ley del logaritmo iterado CSÖRGO)

  Si $F(T) < 1$, entonces

  $$\lim_{n\to\infty} sup \left(\frac{n}{2\log n}\right)^{1/2} \underset{0 \le t \le T}{sup} |F_n^{KM}(t) - F(T)| \le \frac{1}{1 - H(T)}$$

- **Propiedad 6** (Representación casi segura, (LO & SINGH, 1986))

  Bajo la hipótesis de que $F(.)$ y $G(.)$ son continuas, se puede escribir, para todo $t \le T < b_H$:

  $F_n^{KM}(t) + F(t) = n^{-1} \sum_{i=1}^{n} \xi(Z_i, \delta_i, t) + r_n(t)$,

  Donde

  $\xi(Z, \delta, t) = (1 - F(t)) \left[g(Z \wedge t) + \frac{1}{1-H(Z)} 1_{\{Z \le t, \delta = 1\}} \right]$,

  Siendo

  $$g(t) = \int_0^t \left(1 - H(v)\right)^{-2} d(1 - H^1(v))$$

$y \sup_{0 \le t \le T} |r_n(t)| = O\left(\left(\frac{\log n}{n}\right)^{3/4}\right) c.s.$ Además, para todo $a \ge 1$, $\sup_{0 \le t \le T} \mathbb{E}|r_n(t)|^a = o\left(\left(\frac{\log n}{n}\right)^{3a/4}\right)$.

- **Propiedad 7** (Consistencia de la integral Kaplan-Meier, (STUTE & WANG, Astorng law under random censorship, 1993))

  Supongamos que las funciones $F(.)\ y\ G(.)$ no tiene saltos en común, y sea A el conjunto de todos los valores discretos de $H(.)$. Entonces, con la probabilidad uno y en media,

$$\lim_{n\to\infty} \int \varphi(v) dF_n^{KM}(v) = \int_{\{s \notin A, s < b_H\}} \varphi(v) dF(v) + \sum_{a_i \in A} \varphi(a_i) F\{a_i\}$$

$$= \int_{\{s < b_H\}} \varphi(v) dF(v) + 1_{\{b_H \in A\}} \varphi(b_H) F\{b_H\}$$

  Donde $F\{a\} = F(a) + F(a^-)$.

  En particular, si $F(.)$ es continua y $b_F = b_H$, entonces la expresión anterior se reduce a

$$\lim_{n\to\infty} \int \varphi(v) dF_n^{KM}(v) = \int \varphi(v) dF(v)$$

- **Propiedad 8** (Normalidad asintótica de la integral Kaplan-Meier, (STUTE, The central limit theorem under random censorship, 1995))

  Bajo ciertas condiciones (bastante generales) de integrabilidad de la función $\varphi(.)$, se verifica que

$$\sqrt{n} \int \varphi(v) d\left(F_n^{KM}(v) - F(v)\right) \xrightarrow{d} N(0, \sigma^2)$$

  Siendo

$$\sigma^2 = Var[\varphi(Z)\gamma_0(Z)\delta + \gamma_1(Z)(1-\delta) - \gamma_2(Z)]$$

Donde $\gamma_0(.), \gamma_1(.)\ y\ \gamma_2(.)$ tienen una expresión complicada, pero en el caso en que las funciones $F(.) y\ G(.)$ sean continuas, sus expresiones se reducen a

$$\gamma_0(t) = \frac{1}{1 - G(t)},$$

$$\gamma_1(t) = \frac{1}{1 - H(t)} \int_t^{bH} \varphi(v) dF(v),$$

$$\gamma_2(t) = \int_t^{bH} \varphi(v) \left[\int_t^{bH} 1_{\{w<t, w<v\}} \frac{1 - F(w)}{\left(1 - H(w)\right)^2} dG(w) \right] dF(v)$$

- **Propiedad 9** (Representación de la integral Kaplan-Meier como suma de variables independientes (STUTE, The central limit theorem under random censorship, 1995))

  Bajo las mismas condiciones de integrabilidad de la función $\varphi(.)$ que en la propiedad anterior, tenemos que

$$\int \varphi \mathrm{dF}_\mathrm{n}^{\mathrm{KM}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \varphi(\mathrm{Z_i}) \gamma_0(\mathrm{Z_i}) \delta_\mathrm{i} + \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \gamma_1(\mathrm{Z_i})(1 - \delta_\mathrm{i}) \ - \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \gamma_2(\mathrm{Z_i}) + \mathrm{R_n}$$

  Donde

$$R_n = op(n^{-\frac{1}{2}})$$

- **Propiedad 10** (Consistencia uniforme de los procesos de integrales Kaplan- Meier, (BAE & KIM, 2003))

  Si $\mathcal{F} \subset \mathcal{L}_1(F) \coloneqq \{\varphi: \int |\varphi| dF < \infty\}$ tiene una entropía de recubrimiento, $F(.) y\ G(.)$ no tiene saltos en común, entonces

$$\sup_{\varphi \in \mathcal{F}} \left| \int \varphi d(F_n^{KM} - \tilde{F}) \right| \longrightarrow 0$$

  Donde

$$\tilde{F}(t) = F(t) 1_{(t<b_H)} + \left\{ F(b_H^-) + 1_{\{b_H \in A\}} (F(b_H) - F(b_H^-)) \right\} 1_{\{y \geq b_H\}}$$

En probabilidad y varianza.

- **Propiedad 11** (Teorema central del límite uniforme de los procesos de integrales Kaplan-Meier, (BAE & KIM, 2003))

  Sea $\mathcal{F}$ la clase de funciones reales medibles definidos en $\mathbb{R}$. Sea $(\mathcal{F}, d)$ un espacio métrico, con $N_{[]}(u.\mathcal{F}, d)$ el menor n para el que existe $\{f_{0,\delta}^{l}, f_{0,\delta}^{u}, \ldots., f_{0,\delta}^{l}, f_{0,\delta}^{u}\}$, tal que, para todo $\in \mathcal{F}$ existe algún $0 \leq i \leq n$ para el que $f_{i,\delta}^{l} \leq f \leq f_{i,\delta}^{u}$ y $d\left(f_{i,\delta}^{l}, f_{i,\delta}^{u}\right) < \delta$ y sea $J(\delta) \coloneqq \int_0^{\delta}\left[\log N_{[]}(u, \mathcal{F}, d)\right]^{1/2} du$ $para\ 0 < \delta \leq 1.$

  La integral asociada de la entropía de recubrimiento $\log N_{[]}(u, \mathcal{F}, d)$. Si $J(1) < \infty$ y $\mathcal{F}$ tiene la propiedad puntual minimal, entonces el proceso $U_n(\varphi) = \sqrt{n}\int \varphi d(F_n^{KM} - \tilde{F})$ converge como elemento de $B(\mathcal{F})$, el espacio de las funciones acotadas de $\mathcal{F}$, a un proceso gaussiano $W(\varphi)$ con media $\mathbb{E}\big(W(\varphi)\big) = 0$ y función de covarianzas $cov\big(W(\varphi_1), W(\varphi_2)\big) = Cov\big(\xi(\varphi_1), \xi(\varphi_2)\big)$, donde $\xi(\varphi) = \varphi(Z)\gamma(Z)\delta - \int \varphi d\tilde{F} + \gamma_1(Z)(1-\delta) - \gamma_2(Z)$ y $\gamma(t) = \exp\left(\int_{-\infty}^{t-} \frac{dH^0(v)}{1-H(t)}\right)$ $con\ H^0(t) = P(Z \leq t, \delta = 0)$, $\gamma_1(t) = \frac{1}{1-H(t)}\int_t^{b_H} \varphi(v)\gamma(v) dH^1(v)$,

  $$\gamma_2(t) = \iint 1_{\{w<t, w<v\}} \frac{\varphi(v)\gamma(v)}{(1-H(w))^2} dH^0(w) dH^1(v).$$

- **Propiedad 12** (Consistencia de los estimadores semiparamétricos, (DIKTA, On semiparametric random censorship models, 1998))

Sea $0 < T < \infty$ con $H(T) < 1\ y\ \Theta$ un subconjunto y conexo de $\mathbb{R}^k$. Supongamos que $\hat{\theta}_{0,MV} \in \Theta$ es una solución medible de la ecuación $Grad\big(l_n(\theta)\big) = (D_1 \ln(\theta), \ldots., D_k \ln(\theta)) = 0$, siendo $D_r \ln(\theta_0) = [\partial \ln(\theta)/\partial\theta_r]\,|_{\theta=\theta_0}$ y $\ln(\theta)$ la función de logverosimilitud normalizada de los datos, tal que $\hat{\theta}_{0,MV} \rightarrow \theta_0$ con probabilidad uno. Sea $p(.,\theta)$ también una función con derivadas parciales continuas con respecto a θ para toda $\theta \in \Theta\ y\ t \geq 0$, donde $D_r p(.,\theta)$ es medible para todo $\theta \in \Theta$ y existe un entorno $V(\theta_0) \in \Theta\ de\ \theta_0$ y una función medible M tal que $|D_r p(t,\theta)| \leq M(t) y\ \mathbb{E}[M(X)] < \infty$ para todo $\theta \in V(\theta_0), t \geq 0, y\ 1 \leq r \leq k$. Entonces, con la probabilidad uno, cuando $n \rightarrow \infty$, $\sup_{0 \leq t \leq T} |\Lambda_n^D(t) - \Lambda_F(t)| \rightarrow 0$, $\sup_{0 \leq t \leq T} |F_n^D(t) - F(t)| \rightarrow 0$

- **Propiedad 13** (Consistencia de las integrales semiparamétricas $\int \boldsymbol{\varphi dF_n^D}$, (DIKTA, The strong law under semiparametric random censorship models , 2000))

  Supongamos que $H(.)$ es continua, $\hat{\theta}_{0,MV}$ es medible y tiende a θ_0 con probabilidad 1, que para cada $\varepsilon > 0$ existe un entorno $V(\varepsilon, \theta_0) \subset \Theta$ de θ_0 tal que para todo $\theta \in V$

  $\sup_{x \geq 0} |p(x,\theta) - p(x,\theta_0)| < \varepsilon.$

  Si además

  $$\int_0^{b_H} \frac{|\varphi(t)|}{p(t,\theta_0)(1-H(t))^{\varepsilon}} dF(t) < \infty$$

  Para algún $\varepsilon > 0$, entonces

$\lim_{n\to\infty} \int_0^{\infty} \varphi(t) dF_n^D(t) = \int_0^{b_H} \varphi(t) dF(t)$ con probabilidad uno.

Observación:

Si la función $\varphi(x) = 1_{\{0\leq x\leq t\}}$ verifica las condiciones de la propiedad anterior, y $t \leq b_H$, entonces se obtiene la convergencia puntual del estimador $F_n^D(.)$:

$$F_n^D(t) \to F(t)\ en\ probabilidad$$

Si además la función $\varphi(x) = 1_{\{0\leq x\leq t\}}$ verifica las condiciones de la propiedad anterior para todo $t \leq b_H$, entonces se obtiene la consistencia fuerte del estimador $F_n^D(.)$:

$$\sup_{0 \leq t < b_H} |F_n^D(t) - F(t)| \to 0, c.s.$$

- **Propiedad 14** (Representación del estimador semiparamétrico de la razón de fallo acumulativa $\Lambda_c^D(.)$**,** (DIKTA, Weak representation of the cumulative hazard function under semiparametric random censorship models, 2001)).

  Bajo ciertas condiciones sobre el estimador $\hat{\theta}_{0,MV}$, la función $p(.,\theta)$, y suponiendo que $H(.)$ es continua con $1 \leq r_n < n$, entonces

  $$\Lambda_n^D(t) = \int_0^t \frac{dH_n^1(v)}{1-H(v)} + \int_0^t \frac{H_n(v^-)-H(v)}{(1-H(v))^2} dH^1(v) + R_n^0(t) + R_n^1(1),$$

  Donde:

  $$\sup_{0 \leq t \leq Z_{[r_n]}} |R_n^0(t)| = Op\left(\frac{1}{n-r_n}\right) \text{ y}$$

  $$\sup_{0 \leq t \leq Z_{[r_n]}} |R_n^0(t)| = Op\left(n^{1/2} \sum_{i=1}^{r_n} \frac{1}{n-i+1}\right)$$

- **Propiedad 15** (Acotación uniforme del estimador semiparamétrico de la razón de fallo acumulativa $\Lambda_{\boldsymbol{n}}^{\boldsymbol{D}}(.)$**,** (DIKTA, Weak representation of the cumulative hazard function under semiparametric random censorship models, 2001).

  Bajo ciertas condiciones sobre el estimador $\hat{\theta}_{0,MV}$, la función $p(t,\theta)$, y suponiendo que $H(.)$ es continua con $1 \leq r_n < n$, entonces

  $$\sup_{0 \leq t \leq Z_{[r_n]}} |\Lambda_n^D(t) - \Lambda_F(t)| = Op(1)$$

Estimación de la función de densidad

El estimador basado en el estimador de Kaplan – Meier.

El estimador tipo núcleo es el estimador no paramétrico de la función de densidad más estudiado y utilizado en los últimos años, y es el método en el que nos centraremos para estimar la densidad en presencia de censura aleatoria por la derecha. Desde los trabajos de Rosenblatt (1956) y Parzen (1962), los estimadores de tipo núcleo de la densidad han sido quizás los más populares, aunque los primeros trabajos sobre la versión para datos censurados por la derecha aparecieron hasta 1980 (BLUM & SUSARLA, 1980). En general, los estimadores tipo núcleo de la función de densidad son de la forma:

$$\hat{f}_h(t) = h^{-1} \int K\left(\frac{t-v}{h}\right) d\hat{F}(v) = \int K_h(t-v) d\hat{F}(v) = \left(K_h + \hat{F}\right)(t) \quad (2.5)$$

Donde $K_h(.) = h^{-1}K(./h)$ es la función núcleo reescalada, $h > 0$ es el parámetro de suavización o ventana, el símbolo * denota la convolución, y $\hat{F}(.)$ es un estimador de la función de distribución, en nuestro caso, el estimador de Kaplan- Meier. Las condiciones que normalmente ha de verificar la función núcleo es que sea positiva, simétrica y que $\int K(v)dv = 1$. Por lo tanto, en general, ha de ser una función de densidad, pues así $\hat{f}_h(.)$ también lo será ya que hereda todas las propiedades analíticas del núcleo.

Observación:

Si $F(.)\ y\ G(.)$ son dos funciones de distribución, entonces se define la convolución $F * G(.)$ como la función

$$H(t) = \int_{-\infty}^{\infty} G(t-v)dF(v) \quad \forall t \in \mathbb{R}$$

La operación convolución verifica las propiedades conmutativa, asociativa y distributiva respecto a las suma. Si $G(.)$ es absolutamente continua con densidad $g(.)$, entonces $H(.) = F * G(.)$ es absolutamente continua con densidad $h(.)$ dada por

$$h(t) = \int_{-\infty}^{\infty} g(t-v)dF(v) \quad \forall t \in \mathbb{R}$$

Si además $F(.)$ es absolutamente continua con densidad $f(.)$ entonces

$$h(t) = \int_{-\infty}^{\infty} g(t-v)f(v)dv \quad \forall t \in \mathbb{R}$$

La importancia de la convolución de funciones de distribución viene dada por el hecho de que si $X\, e\, Y$ son dos variables aleatorias independientes con funciones de distribución $F(.) y\, G(.)$ respectivamente, entonces la función de distribución de la variable suma $S = X + Y$ es

$$H(t) = \int_{-\infty}^{\infty} G(t-v)dF(v) \quad \forall t \in \mathbb{R}$$

Como ya se comento antes, no fue hasta principios de los años 80 cuando se publicaron las primeras propiedades del estimador tipo núcleo de la función de densidad con datos censurados por la derecha. Fodes, Rehtö y Winter obtuvieron, tomando como $\hat{F}(.)$ en la expresión (1.6), el estimador de Kaplan – Meier, la convergencia fuerte para le estimador núcleo de la densidad $f_n^{KM}(.)$, que se reduce al estimador usual tipo núcleo de Parzen (1962) en el caso de no censura (puesto que $F_n^{KM}(.)$, en tal caso, so convierte en la función de distribución empírica usual, $F_n(.)$).

En la estimación tipo núcleo de la función de densidad (con o sin censura, es bien conocido que la elección de la función núcleo tiene una importancia menor, y son comunes en la práctica los núcleos dados en la TABLA 1.

Sin embargo, la bondad de cualquier procedimiento estadístico que implique suavización, y, en concreto, la estimación tipo núcleo de la función de densidad, dependen de manera crucial de la elección del parámetro de suavización o ventana, ya que regula el grado de suavización del estimador. Ventanas

demasiado pequeñas producen estimaciones con mucha variabilidad (infrasuavizacion), mientras que ventanas grandes dan lugar a estimaciones sesgadas (sobresuavizacion).

2.2.4 Estrés

La palabra estrés significa diferentes cosas para diferentes personas, estas definiciones se pueden clasificar en función de la conceptualización del estrés como estímulo, repuesta, percepción o transacción. Las tres conceptualizaciones más conocidas del estrés son las siguientes:

Como un conjunto de estímulos: Existen ciertos estímulos en el ambiente que nos producen tensión y/o se perciben como amenazas peligrosas (CANNON, 1929).

Como una respuesta: Enfocado más en cómo reaccionan las personas antes los estímulos (SELYE, 1973). Se puede distinguir dos componentes: el psicológico y el filosófico.

Como un proceso: Que incorpora tanto estresores como las respuestas a los mismos y además añade la interacción entre la persona y el ambiente (LAZARUS & FOLKMAN, 1987). El determinante crítico del estrés es cómo la persona percibe y responde a diferentes acontecimientos.

El estrés puede definirse como el estado que se manifiesta por un síndrome específico consistente en todos los cambios inespecíficos inducidos dentro de un sistema biológico pero sin una causa particular (SELYE, 1973).

También podemos definir el estrés psicológico como el resultado de una relación particular entre un individuo y el entorno que es evaluado por este como amenazante ó desbordante de sus recursos y que pone en peligro su bienestar (LAZARUS & FOLKMAN, 1987).

Sin embargo el estrés no es malo en todo sentido es más bien una respuesta a los estímulos, por lo que influye mucho la forma de percepción que cada individuo tenga sobre los mismos, entonces podemos definir el estrés como una repuesta que se manifiesta en los individuos al percibir en el entorno situaciones que son interpretadas por este como amenazante.

El estrés como estimulo

El estrés en este modelo puede definirse como un estimulo de ambiente en el individuo, asumiendo que estos pueden causas perturbaciones o alterar el comportamiento normal de organismo. Dado que constantemente se tiene estímulos y nosotros contamos con recursos para afrontar a estos estímulos, a veces se cuenta con los recursos necesarios (conocimientos, materiales, etc.) por lo que estos estímulos son fácilmente afrontados, sin embargo cuando estos estímulos sobrepasan a nuestros recursos, todo nuestro organismo se ve alterando ante la amenaza.

El estrés como proceso

El estrés también puede ser entendido en términos de las interpretaciones cognitivas que la persona hace sobre la capacidad estresora de los eventos. Desde este punto de vista las personas evalúan las situaciones y/o estímulos del ambiente, en esta parte se puede distinguir tres tipos de evaluación (LAZARUS & FOLKMAN, 1987):

Primero: La persona valora el significado de los que está ocurriendo.

Segundo: La persona valora sus propios recursos para afrontar la situación.

Tercero: Procesos de retroalimentación que se desarrolla en la interacción del individuo con las demandas externas e internas.

Fuentes de estrés

Las fuentes de estrés básicas son: el entorno, el propio cuerpo y los propios pensamientos.

En estrés puede provenir del ambiente, dado que este bombardea al sujeto con constantes demandas de adaptación, como: ruido, aglomeraciones, relaciones interpersonales o los horarios rígidos.

La segunda fuente de estrés proviene de nuestro propio cuerpo, como las fuentes de estrés pueden afectar y producir efectos negativos y traumáticos en las persona.

La tercera y última fuente son nuestros pensamientos, en la que interpretamos y clasificamos nuestras experiencias, y la forma de percibir el futuro entorno a ellas.

Las demandas psicosociales a las que se enfrentan las personas en interacción con los recursos de que disponen para acometerlas, pueden originar una serie de consecuencias fisiológicas, cognitivas y motoras sobre su estado de salud (KIVIMÄKI, VAHTERA, ELOVAINIO, LILLRANK, & KEVIN, 2002).

Síndrome de Burnout

Es un tipo de estrés laboral, que en los últimos años ha acaparado gran medida la atención de investigadores de estrés laboral. Se estima que siete de cada diez trabajadores se siente quemados por su trabajo (FERNÁNDEZ MARTÍNEZ, 2009). Gran parte de estos estudios se tratan sobre los antecedentes, consecuentes, facilitadores, factores protectores y una gran variedad de variables que ayudan a delimitar y a definir mejor este síndrome que es la respuesta del individuo al estrés laboral.

Este síndrome fue descrito por primera vez por la psiquiatra estadounidense Herbert Freudenberger en 1974 (LÓPEZ, ZEGARRA, & CUBA, 2006). Observando que la mayoría de los profesionales que trabajaban junto a ella sufrían de un desgaste o perdida de energía y vitalidad e incluso sensibilidad con el tiempo. Además Freudenberger describió que estas personas

con el tiempo se volvían incluso agresivas hacia los pacientes, culpándolos muchas veces por sus problemas personales. Más adelante en el año de 1976 una psicóloga social Cristina Maslach estudia las respuestas emocionales en empleados al servicio de pacientes, sentando posteriormente las bases conceptuales y empíricas de este síndrome (MASLACH & JACKSON, 1981).

Según Maslach el burnout puede ser definido como el síndrome de extenuación emocional, despersonalización y falta de logro personal en el trabajo. Es un trastorno adaptativo crónico asociado al inadecuado afrontamiento de las demandas psicológicas en el trabajo, que daña la calidad de vida de la persona, y disminuye la calidad de su trabajo.

Farber define el burnout de la siguiente manera : “El burnout es un síndrome relacionado con el trabajo. Surge por la percepción del sujeto de una discrepancia entre los esfuerzos y lo conseguido. Sucede con frecuencia en los profesionales que trabajan cara a cara con clientes necesitados o problemáticos. Se caracteriza por un agotamiento emocional, falta de energía, distanciamiento y cinismo hacia los destinatarios, sentimiento de incompetencia, deterioro de autoconcepto profesional, actitudes de rechazo hacia el trabajo y por diversos síntomas psicológicos como irritabilidad, tristeza y baja autoestima” (FARBER, 1983).

Con respecto al Síndrome de pueden distinguir dos tipos de perspectivas:

Perspectiva clínica: La perspectiva clínica entiende el síndrome de quemarse como un estado al que llega el sujeto como consecuencia del estrés. Hace alusión a la experiencia de agotamiento, decepción y pérdida de interés por la actividad con otras personas. Se combinan fatiga emocional, física y mental, sentimientos de impotencia e inutilidad, sensaciones de sentirse atrapado, falta de entusiasmo en las actividades diarias y la vida general, y baja autoestima. También puede ser entendido como un trauma narcisista que conlleva una disminución en la autoestima de los sujetos.

Perspectiva psicosocial: La perspectiva psicosocial apunta hacia la consideración como un proceso que se desarrolla por la interacción de características del entorno y características personales. El síndrome de quemarse como proceso asume una secuencia de etapas o frases diferentes con sintomatología, a su vez, diferenciada. Desde esta perspectiva el síndrome de quemarse por la rutina diaria no debe identificarse como estrés psicológico, sino que debe ser entendido como una respuesta a fuentes de estrés crónico que surge de las relaciones sociales entre proveedores de los servicios y receptores de los mismos. Es un tipo particular de mecanismos de afrontamiento y autoprotección frente al estrés generado en la relación con otras personas, y en la relación profesional – organización.

Factores característicos del Síndrome de Burnout

Presentar síntomas físicos de estrés psicofisiológico, como cansancio hasta el agotamiento, malestar general, junto con técnicas paliativas reductoras de la ansiedad residual.

Alteraciones de conducta (Conducta anormal del modelo asistencial o despersonalización de la relación con el paciente).

Síntomas disfóricos y de agotamiento emocional.

Menor rendimiento laboral, desmotivación y retirada organizacional.

Es un síndrome clínico – laboral que se produce por una inadecuada adaptación al trabajo, aunque se presente en individuos considerados presuntamente normales.

Componentes del Síndrome de Burnout

Según Maslach son tres los principales componentes o indicadores relacionados con el síndrome de Burnout. Estos también estas relacionados empíricamente entre si.

Cansancio emocional: Es una situación de agotamiento de la energía o de los recursos emocionales propios, una experiencia de estar emocionalmente agotado debido al contacto diario y sostenido con personas con las que hay que interactuar. Aparece el desgaste, la fatiga y manifestaciones físicas y psíquicas en representación del vaciamiento de los recursos emocionales y personales, experimentándose una sensación de

que no tener nada más que ofrecer profesionalmente. Incluye disminución y pérdida de recursos emocionales.

Despersonalización: En el sentido de deshumanización: incluye indiferencia, un sentimiento de distancia emocional y el desarrollo de actitudes negativas de insensibilidad y cinismo hacia los receptores del servicio prestado. Garden describe cuatro factores dentro del ítem de despersonalización: distanciamiento, hostilidad, despreocupación por los demás, rechazo y falta de interés por los otros.

Desarrollo de actitudes, cinismo y sentimientos negativos hacia las personas destinatarias. Los sujetos presentan un incremento en la irritabilidad, con pérdida de motivación, reacciones de distanciamiento y hostilidad hacia los pacientes y compañeros de trabajo.

Falta de realización personal (bajos sentimientos de realización personal): Incluye vivencia de insuficiencia personal, sentimientos de fracaso y baja autoestima con una tendencia a evaluar su desempeño realizado de manera negativa (y con auto- reproches por no haber alcanzado los objetivos propuestos.)

Faces evolutivas associadas al Síndrome de Burnout

Como consecuencia de los estresores laborales, los trabajadores desarrollan sentimientos de agotamiento emocional que posteriormente dan lugar a la aparición de una actitud

despersonalizada hacia las personas que deben atender, y como consecuencia de ello pierden el compromiso personal y disminuye su realización personal en el trabajo. El agotamiento emocional sería la dimensión fundamental del Burnout, a la que seguiría despersonalización y, posteriormente, la reducida realización personal. El agotamiento emocional, por tanto, sería la dimensión que ocasionaría la baja realización personal, estando este proceso mediatizado por despersonalización. Este proceso vendrá determinado por las interacciones que cada dimensión mantenga con los diversos componentes organizacionales: competencia (fomento de habilidades y afrontamiento efectivo), autonomía y participación en la toma de decisiones, colegiación (apoyo del supervisor y de los compañeros), y cooperación con el cliente. Según este autor, el agotamiento emocional sería lo que sentiría un trabajador con Burnout y, por tanto, la dimensión que pondría en marcha el síndrome, siendo la baja realización personal la que daría lugar a las consecuencias observables del Burnout como el absentismo, la rotación de puestos, el abandono del trabajo, etc.), mediatizadas por la despersonalización

Variables predisponenetes

Las variables estudiadas como posibles desencadenantes o facilitadores del burnout provienen de fuentes diversas que en general suelen estar relacionadas con las situaciones sociales y

demográficas del sujeto, con el ambiente y condiciones en que desarrolla su trabajo y con su propia disposición y características personales.

Existen una serie de factores de riesgo que influyen en su aparición:

La edad. Muchas investigaciones coinciden en que a mayor edad menos propensas son las personas a sufrir en síndrome de burnout, dado que las estrategias de afrontamiento de estas personas se han desarrollado mejor con el tiempo.

El sexo. La mujeres tienden a demostrar más sus sentimientos con respecto a las perturbaciones que los hombres, lo que las hace más vulnerables.

El estado civil. Las personas que tienen alguien con quien compartir estas experiencias adquiridas son más propensas a sufrir el síndrome.

El horario laboral. Cuando este se presenta en exceso puede ocasionar problemas.

Antigüedad profesional. Las personas más antiguas tienden a tener una mayor despersonalización, y los de menor antigüedad tiende a tener un indicador alto de falta de realización personal.

La sobrecarga laboral. Produce una diminución de la calidad de atención.

Salario.

Evaluaciones.

Apoyo social. Ayuda a las persona a superarse profesionalmente, sentirse útil, valorado y querido.

Conflictos entre los valores personales y organizacionales.

Burnout en estudiantes universitarios

El burnout en los estudiantes universitarios puede ser definido como “Un conjunto de síntomas psicológicos que ocurren debido al estrés académico crónico y a las cargar del curso, manifestando por agotamiento en los estudiantes, desanimo respecto a las tareas de aprendizaje y eficacia profesional reducida” (GAN, SHANG, & ZHANG, 2007).

La mayoría de estudios sobre estrés en universitarios demuestra que existen índices elevados en los estudiantes de bajos semestres o inicios de carrera. Los estresores más importantes son la preocupación por el desempeño, el proceso de adaptación al ambiente universitario, la exigencias de los estudios, las notas finales, el excesivo trabajo para casa, los exámenes y el estudiar para ellos, la incertidumbre hacia el futuro (oferta laboral). Además de la falta de relaciones sociales verdaderas. Las universidades preocupadas por ofrecer enseñanza de calidad deben tener en cuenta las variables relacionadas con el proceso de enseñanza - aprendizaje (SALANOVA, GRAU, & MARTÍNEZ, 2005).

El interés que despierta el burnout facilita el estudio de este síndrome en diferentes campos, no solamente en profesionales

que interactúan con personas diariamente en su trabajo, sino también en toda persona que sufra de condiciones de exigencia elevada, critica continua y no pueda administrar la exigencia que le es impuesta por la institución (FERNÁNDEZ MARTÍNEZ, 2009).

En el nuestro ambiente podemos percibir como estudiantes que los primeros días de clases tiene un entusiasmo, van decayendo poco a poco con el transcurrir del tiempo, y que estos estudiantes tienen pocas expectativas sobre acabar sus estudios con éxitos y también están poco preparados para enfrentarse al mundo laboral.

Se ha realizado muchos estudios sobre el burnout en profesionales, sin embargo son pocos los estudios realizados en estudiantes, siento este de gran importancia puesto que se trata de futuros profesionales. El periodo de estudio universitario es percibido como un evento que conlleva grandes oportunidades para el desarrollo personal y representa en desarrollo en la etapa final de la adolescencia y en los jóvenes adultos (FERNÁNDEZ MARTÍNEZ, 2009). Los estudiantes que sufren de burnout dejan de ir a clases, no tiene ánimo para ellos y no les importa, la presión del tiempo para presentar trabajos y la disposición de horarios muy ajustados están relacionadas con la disminución del rendimiento.

2.3 DEFINICIÓN DE TÉRMINOS BÁSICOS

Agotamiento físico y psíquico

El agotamiento físico y psíquico se refiere a la falta de energía (física, mental y emocional) de las personas, el cual es producido muchas veces por el contacto con personas, el anteponer sus necesidades a las nuestras, viviendo así una experiencia de no tener nada más que ofrecer a los demás.

Despersonalización

Definida como la deshumanización, esta incluye la indiferencia y distancia emocional con los sentimientos de las personas con las que se tiene contacto. En el Test del síndrome de Burnout se pueden detectar cuatro factores: distanciamiento, hostilidad, despreocupación por los demás, rechazo y falta de interés por los otros.

Edad

La edad la definiéremos como la cantidad de años transcurridos desde el día de nacimiento de la persona hasta la fecha actual (el día de hoy, o en nuestro caso el día de la aplicación de la encuesta).

Escuela profesional

La escuela profesional es el lugar donde tiene lugar la enseñanza de las materias para formar profesionales.

Estrés

"Es el resultado de una relación particular entre el individuo y el entorno que es evaluado por este como amenazante o desbordante de sus recursos y que pone en peligro su bienestar" (LAZARUS & FOLKMAN, 1987)

Estudiante

Toda aquella persona que este matriculado en al menos un curso dentro de la Universidad Nacional del Altiplano

Evento terminal

Es el evento por el cual se deja de estudiar al sujeto, dado que denota la presencia del suceso de interés (REBASA, 2005). Para el caso de nuestra tesis fue la Presencia del Síndrome de Burnout.

Falta de realización personal

Es la tendencia de la persona a considerar con un rendimiento bajo en sus actividades, las personas se sienten infelices consigo mismas y mostrando un declive en el sentimiento de competencia.

Fecha de cierre

Es la fecha límite para a toma de la última observación a las unidades estudiadas, definen el termino del tiempo calendario de la observación.

Fecha de inicio

El la fecha que define el intervalo de tiempo de duración del estudio, esta fecha indica el inicio del tiempo calendario.

Fecha de la última observación

Es la última fecha en la que se registra la última noticia acerca del sujeto de estudio.

Función de núcleo

"Es una función $k(x)$, a partir de la cual se puede establecer un estimador no parametrico de cualquier funcion de densidad $f(x)$ " (ROSENBLATT, 1956)

Lugar de procedencia

Es el lugar de nacimiento de la persona, de manera tal que defina algunos rasgos físicos en ella.

Parámetro de alisado o suavización

Este parámetro también llamado "ancho de ventana" o "ancho de banda", "es un número positivo que cumple con la condición de Parzen" (MARTÍNEZ FALERO, AYUGA TÉLLEZ, & GONZÁLES GARCÍA, 1992) En un parámetro que controla en equilibrio entre el sesgo y la varianza de los datos con los que se trabaja, especificando el tamaño de la zona local.

Sexo

Características físico-biológicas que definen si una persona es varón o mujer.

Síndrome de Burnout

La definición más conocida es la propuesta por Maslach y Jackson “Es un síndrome de agotamiento emocional, despersonalización y baja realización personal, que puede ocurrir entre individuos que interactúan diariamente con otras personas” (MASLACH & JACKSON, 1981).

Dicha enfermedad mental es la respuesta prolongada al estrés repercutiendo muchas veces en consecuencias físicas.

El burnout en estudiantes puede definirse como un conjunto de síntomas psicológicos que ocurren debido al estrés académico crónico y a las cargas del curso, manifestado por agotamiento en los estudiantes, cinismo respecto de sus tareas de aprendizaje y eficiencia profesional reducida (GAN, SHANG, & ZHANG, 2007).

Tiempo calendario

El tiempo calendario de un intervalo de tiempo en él se ejecutara el seguimiento a los pacientes, este intervalo es general dado que tiene un tiempo de inicio del seguimiento establecido y un tiempo límite de seguimiento establecido.

Tiempo de seguimiento

Es el intervalo de tiempo total para cada uno de los estudiantes que se registra hasta la observación del suceso en estudio, para el caso de la tesis la presencia del Síndrome de Burnout.

2.4 OPERACIONALIZACIÓN DE VARIABLES

En la Tabla 2, se muestran las variables indicadores e índices.

TABLA 2

OPERACIONALIZACIÓN DE VARIABLES

TIPO	DIMENSIÓN	INDICADOR	INDICE
INDEPENDIENTE	DEMOGRÁFICA	Sexo	Femenino Masculino
		Edad	Menor a 18 años Entre 18 a 21 años Entre 22 a 25 años Más de 25 años
	SOCIO ECONÓMICA	Tiempo de Estudio	Primer y segundo semestre. Tercer y cuarto semestre. Quinto y sexto semestre. Séptimo y octavo semestre. Noveno y décimo semestre.
		Lugar de Procedencia	Puno Juliaca Ilave Otros
		Gasto Mensual	Menos de 100 al mes Entre 100 y 200 soles al mes Más de 200 soles al mes
		Lugar de Residencia	Si = Vive en Puno No = Vive en otros lugares
		Situación de Convivencia	Solo Con familiares Con amigos Otros
		Propiedad de La Residencia	En habitación alquilada En casa de un familiar Otros
DEPENDIENTE	SÍNDROME DE BURNOUT	Cansancio emocional Despersonalización Realización personal	0= No está presente 1= Si está presente

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1 POBLACIÓN

La población estuvo formada por todos los estudiantes de la Universidad Nacional del Altiplano Puno matriculados en el semestre 2013 – I, los que según la Oficina de Estadística de la Universidad Nacional del Altiplano son en total 17364 estudiantes (Anexo Nº 01).

3.2 MUESTRA

Para la determinación del tamaño muestral se utilizó el método de muestreo estratificado, cuya fórmula es la siguiente:

$$n_0 = \frac{Z^2(\sum_{h=1}^{L} W_h P_h Q_h)}{E^2}$$

Donde:

h: 1,2,3 ... 35

L: Número total de estratos.

n_0: Tamaño muestral general.

Z: Valor de la distribución normal.

W_h: Proporción para cada estrato.

P_h: Proporción de alumnos con síndrome de Burnout de cada estrato.

Q_h: Proporción de alumnos sin síndrome de Burnout de cada estrato.

E: Error muestral.

El cálculo del error muestral se realizó a través de la siguiente fórmula

$$E = e * P_g$$

Donde:

e: Error de la proporción muestral.

P_g: Proporción de alumnos con síndrome de Burnout de la muestra.

El cálculo de la proporción general de la muestra se realizó con la siguiente fórmula:

$$P_g = \sum_{h=1}^{L} W_h P_h$$

Los estratos de nuestra población fueron constituidos por las facultades (Anexo Nº 01).

CÁLCULO DE LA MUESTRA

En la TABLA 3 se muestran las proporciones y ponderaciones de cada estrato.

TABLA 3

PROPORCIONES Y PONDERACIONES DE LA POBLACIÓN ESTUDIANTIL SEGÚN ESTRATO.

ESCUELA PROFESIONAL	TOTAL	Wh	Ph(*)	Qh(*)
Ingeniería Agronómica	347	0,019983875	0,8	0,2
Ingeniería Agroindustrial	295	0,016989173	0,75	0,25
Ingeniería Topográfico y Agrimensura	430	0,024763879	0,72	0,28
Medicina Veterinaria y Zootecnia	718	0,041349919	0,36	0,64
Ingeniería Económica	730	0,042041004	0,56	0,44
Ciencias Contables	919	0,052925593	0,85	0,15
Administración	681	0,039219074	0,46	0,54
Enfermería	539	0,031041235	0,87	0,13
Trabajo Social	546	0,031444368	0,6	0,4
Ingeniería de Minas	689	0,039679797	0,75	0,25
Sociología	503	0,02896798	0,758	0,242
Turismo	501	0,028852799	0,445	0,555
Antropología	340	0,019580742	0,87	0,13
Ciencias de la Comunicación Social	579	0,033344851	0,74	0,26
Arte	386	0,022229901	0,89	0,11
Biología	426	0,024533518	0,98	0,02
Educación Secundaria	821	0,047281732	0,78	0,22
Educación Física	203	0,011690855	0,74	0,26
Educación Primaria	326	0,018774476	0,85	0,15
Educación Inicial	297	0,017104354	0,8	0,2
Ingeniería Estadística e Informática	324	0,018659295	0,83	0,17
Ingeniería Geológica	582	0,033517623	0,79	0,21
Ingeniería Metalúrgica	353	0,020329417	0,84	0,16
Derecho	652	0,037548952	0,85	0,15
Ingeniería Química	280	0,016125317	0,85	0,15
Nutrición Humana	330	0,019004838	0,842	0,158
Odontología	499	0,028737618	0,853	0,147
Ingeniería Agrícola	479	0,02758581	0,75	0,25
Ingeniería civil	804	0,046302695	0,45	0,55
Arquitectura y Urbanismo	595	0,034266298	0,856	0,144
Ciencias Físico – Matemáticas	167	0,0096176	0,8512	0,1488
Medicina Humana	420	0,024187975	0,45	0,55
Ingeniería Mecánica Eléctrica	631	0,036339553	0,74	0,26
Ingeniería Electrónica	428	0,024648698	0,75	0,25
Ingeniería de Sistemas	544	0,031329187	0,788	0,212
TOTAL	17364	1		

Fuente: Encuesta piloto aplicada por la investigadora.

(*) Proporciones calculadas a partir de la encuesta piloto aplicada a 20% de la muestra total, 102 estudiantes de la Universidad Nacional del Altiplano Puno.

La muestra determinada fue de 511 estudiantes de la Universidad Nacional del Altiplano (Ver Anexo Nº11), a los cuales se les realizó el seguimiento respectivo.

Los tamaños muestrales para los estratos fueron calculados a través de la siguiente fórmula:

$$n_h = n_0 * W_h$$

Reemplazando se obtiene los siguientes tamaños muestrales por escuela profesional mostradas en la TABLA 4

TABLA 4

TAMAÑOS MUESTRALES POR ESTRATO

ESCUELA PROFESIONAL	TOTAL	TAMAÑO MUESTRAL
Ingeniería Agronómica	347	10
Ingeniería Agroindustrial	295	9
Ingeniería Topográfico y Agrimensura	430	13
Medicina Veterinaria y Zootecnia	718	21
Ingeniería Económica	730	21
Ciencias Contables	919	27
Administración	681	20
Enfermería	539	16
Trabajo Social	546	16
Ingeniería de Minas	689	20
Sociología	503	15
Turismo	501	15
Antropología	340	10
Ciencias de la Comunicación Social	579	17
Arte	386	11
Biología	426	13
Educación Secundaria	821	24
Educación Física	203	6
Educación Primaria	326	10
Educación Inicial	297	9
Ingeniería Estadística e Informática	324	9
Ingeniería Geológica	582	17
Ingeniería Metalúrgica	353	10
Derecho	652	19
Ingeniería Química	280	8
Nutrición Humana	330	10
Odontología	499	15
Ingeniería Agrícola	479	14
Ingeniería civil	804	24
Arquitectura y Urbanismo	595	17
Ciencias Físico – Matemáticas	167	5
Medicina Humana	420	12
Ingeniería Mecánica Eléctrica	631	19
Ingeniería Electrónica	428	13
Ingeniería de Sistemas	544	16
TOTAL	**17364**	**511**

Fuente: La investigadora

3.3 MÉTODOS DE RECOPILACIÓN DE DATOS

RECOLECCIÓN DE VARIABLES DE INTERES

Para recolectar las variables de interés como son el sexo, la edad, tiempo de estudio, lugar de procedencia, gasto mensual, lugar de residencia, situación de convivencia, propiedad de la residencia, se procedió a la utilización de una encuesta elaborada para esta investigación (Anexo Nº 02).

PARA EL SÍNDROME DE BURNOUT

Para recolectar la información acerca de la presencia del Síndrome Burnout en los estudiantes se utilizó el test Maslach Burnout Inventory, que contiene 22 ítems (Anexo Nº 03) distribuidos de la siguiente forma para cada una de los componentes de síndrome de Burnout:

TABLA 5

COMPONENTES DEL SÍNDROME DE BURNOUT E ÍTEMS CORRESPONDIENTES SEGÚN EL MASLACH BURNOUT INVENTORY

COMPONENTE DEL SÍNDROME DE BURNOUT	ITEMS CORRESPONDIENTES
Cansancio emocional	1, 2, 3, 6, 8, 13, 14, 16 y 20
Despersonalización	5, 10, 11, 15 y 22
Realización personal	4, 7, 9, 12, 17, 18, 19 y 21

3.4 MÉTODOS DE TRATAMIENTO DE DATOS

3.4.1 Recodificación de las variables y determinación de la Presencia del Síndrome de Burnout.

- Recolección de datos a través de los instrumentos de medición (Anexo N°02 y Anexo N°03). Realizado durante 10 semanas.
- Procesamiento de los datos y obtención de la matriz de datos.
- Recodificacion de las variables y obtención del las escalas para cada componte del síndrome de Burnout según la siguiente puntuación.

TABLA 6

PUNTUACIÓN POR ESCALA DE LOS COMPONENTES DEL SÍNDROME DE BURNOUT

PUNTUACIÓN ESCALA	ALTA	MEDIA	BAJA
Cansancio emocional	≥ 27	$19 - 26$	≤ 18
Despersonalización	≥ 10	$6 - 9$	≤ 5
Realización personal	≥ 40	$34 - 39$	≤ 33

- Obtención del indicador de la presencia del síndrome de Burnout, de acuerdo a las puntuación de los componentes, Puntuaciones altas en Cansancio emocional y Despersonalización, y una puntuación baja en Realización personal, denota presencia del síndrome, en otro caso no se presenta el síndrome, se codifico la variable síndrome de Burnout como se describe a continuación:

$$y = \begin{cases} 1\ si\ presenta\ el\ Sindrome \\ 0\ \ no\ presenta\ el\ Sindrome \end{cases}$$

3.4.2 Estimación de modelos de regresión logística con datos censurados y no censurados

- Estimación de la función de distribución para el Síndrome de Burnout, se utilizó las funciones de regresión no parametrica, se utilizó el paquete Kerdiest del R Project.
- Prueba de censura de datos a través del estadístico Kaplan Meier.
- Estimación del modelo de regresión logística con datos censurados a través del SPSS Versión 19.
- Estimación del modelo de regresión logística sin datos censurados a través del SPSS Versión 19.
- Ambos modelos bajo el siguiente modelo teórico:

$$p_i = \frac{1}{1 + \exp(-\alpha - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \beta_4 x_{4i} - \beta_5 x_{5i} - \beta_6 x_{6i} - \beta_7 x_{7i} - \beta_8 x_{8i})}$$

Donde:

p_i: Representa la presencia del Síndrome de Burnout, tomara dos valores
0= No presenta el Síndrome de Burnout.
1 =Si presenta el Síndrome de Burnout.

a: Es la media general de los datos.

β_1: Representa el coeficiente estimado para la variable Sexo.

β_2: Representa el coeficiente estimado para la variable Edad.

β_3: Representa el coeficiente estimado para la variable Tiempo de estudio.

β_4: Representa el coeficiente estimado para la variable Lugar de procedencia.

β_5: Representa el coeficiente estimado para la variable Gasto mensual.

β_6: Representa el coeficiente estimado para la variable Lugar de residencia.

β_7: Representa el coeficiente estimado para la variable Situación de convivencia.

β_8: Representa el coeficiente estimado para la variable Situación de la residencia.

x_{1i}: Es la observación de la variable Sexo para el i–ésimo sujeto.

x_{2i}: Es la medición de la variable Edad para el i–ésimo sujeto.

x_{3i}: Es la medición de la variable Tiempo de estudio para el i–ésimo sujeto.

x_{4i}: Es la observación de la variable Lugar de procedencia para el i–ésimo sujeto.

x_{5i}: Es la observación de la variable Gasto mensual para el i–ésimo sujeto.

x_{6i}: Es la observación de la variable Lugar de residencia para el i–ésimo sujeto.

x_{7i}: Es la observación de la variable Situación de convivencia para el i–ésimo sujeto.

x_{8i}: Es la observación de la variable Situación de residencia para el i–ésimo sujeto.

3.4.3 Comparación de los modelos de regresión

Para poder realizar la comparación de ambos modelos se utilizó la estimación tipo kernel con parámetros de suavizado mediante la cual se obtuvieron las medidas como el sesgo asintótico, pseudoverosimilitud y error medio integrado; estas medidas se utilizaron con el fin de establecer cuál de ambos modelos estimados explicaba mejor la presencia del síndrome de Burnout, todas las estimación se realizaron mediante el paquete R Project, los pasos fueron:

- Estimación del parámetros de suavizado.
  - Por el método de validación cruzada.
  - Por el método Plug-in, de Ruppert, Shealther y Wand , y el estimador de Polansky y Baker.
  - La estimación de estos parámetros se basan en la estimación de la siguiente función:

$$MISE(\hat{F}_h) = \int_{-\infty}^{+\infty} (\hat{F}_h(x) - F(x))^2 dx$$

    Estos cálculos están implementados en el paquete Kerdiest del R Project.
- Elección del mejor parámetro de suavizado, se realizara mediante el valor calculado del AMISE, el cual se obtiene mediante el paquete sm del R Project.

- Estimación de la función de densidad para el modelo de regresión logística con datos censurados, a través de los estimadores tipo núcleo y el parámetro de suavizado, funciones implementadas en el paquete locpol de R Project.
- Estimación de la función de densidad para el modelo de regresión logístico con datos no censurados a través de los estimadores tipo núcleo y el parámetro de suavizado, funciones implementadas en el paquete locpol de R Project.
- Estimación del sesgo asintótico para ambas funciones de densidad de los modelos implementado en el paquete sm del R Project.
- Estimación de la Pseudoverosimilitud para ambas funciones de densidad de los modelos implementado en el paquete sm del R Project.
- Estimación del Error medio integrado para ambas funciones de densidad de los modelos implementado en el paquete sm del R Project
- Comparación del sesgo asintótico, la pseudoverosimilitud y el error medio integrado de ambas funciones de densidad de los modelos.
- A través de la medidas se concluirá cual de ambos modelos es el que mejor puede interpretar los datos.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 RESULTADOS Y DISCUSIÓN

4.1.1 Características de la muestra de estudio

Antes de realizar la estimación de los modelos de regresión logística con datos censurados y no censurados, se observaron las características principales de la muestra estudiada, las que se parecían a continuación.

TABLA 7

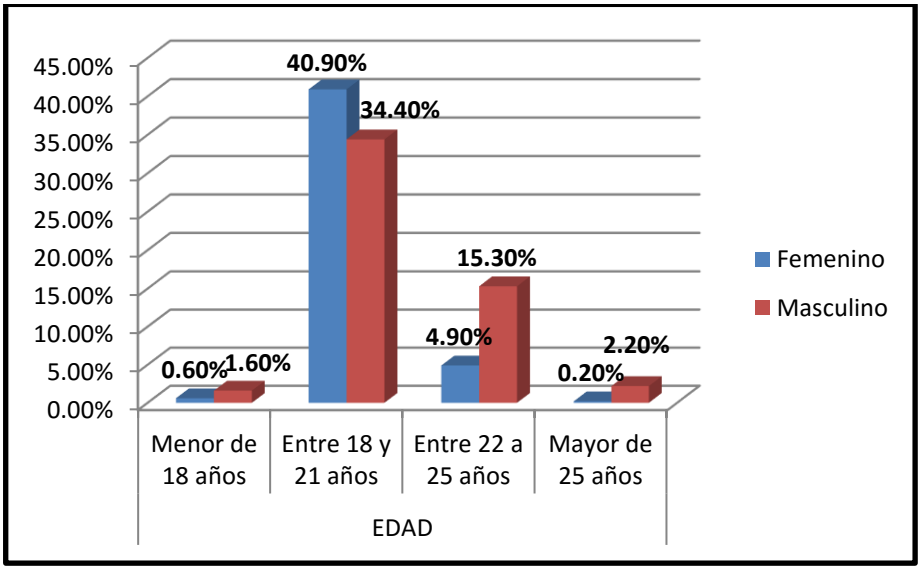
PORCENTAJES Y NÚMERO DE ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 – II SEGÚN SEXO Y EDAD

SEXO		EDAD				Total
		Menor de 18 años	Entre 18 y 21 años	Entre 22 a 25 años	Mayor de 25 años	
Femenino	n	3	209	25	1	238
	%	0,6%	40,9%	4,9%	0,2%	46,6%
Masculino	n	8	176	78	11	273
	%	1,6%	34,4%	15,3%	2,2%	53,4%
Total	n	11	385	103	12	511
	%	2,2%	75,3%	20,2%	2,3%	100,0%

Fuente: Matriz de datos

GRÁFICO 1

PORCENTAJES DE ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013-II SEGÚN SEXO Y EDAD



Fuente: Tabla 6

INTERPRETACIÓN: En la Tabla 7 y el Gráfico 1 se observa que de los 511 estudiantes, que participaron voluntariamente en la investigación el 53.42% (273) fueron de sexo masculino y el 46.58% (238) restante fueron de sexo femenino. Entre los cuales a edad promedio general fue de 20.63 (±2.74) años.

TABLA 8

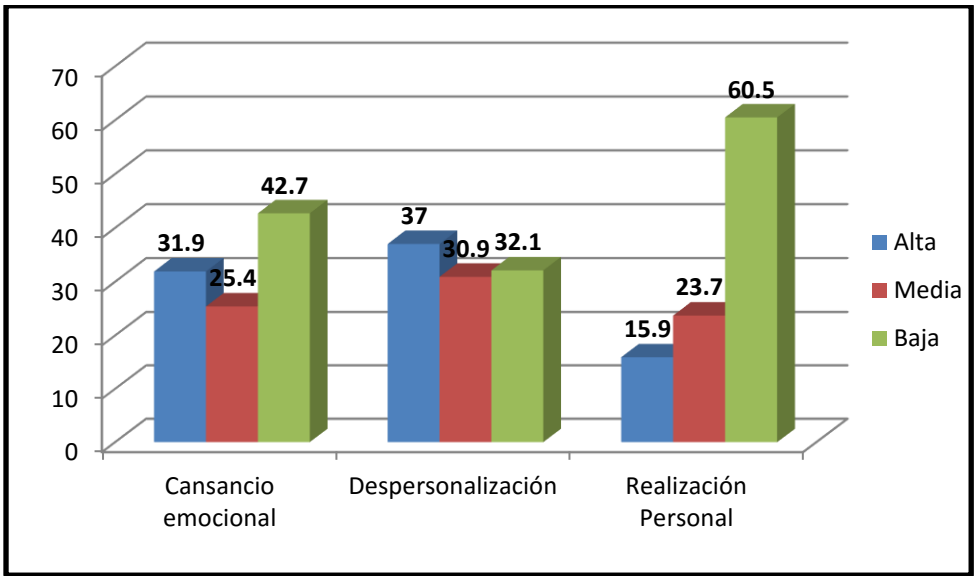
PORCENTAJES Y NÚMERO DE ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 - II SEGÚN PUNTUACIÓN POR COMPONENTE DEL SÍNDROME DE BURNOUT

COMPONENTE		PUNTUACIÓN		
		Alta	Media	Baja
Cansancio emocional	n	163	130	218
	%	31.9	25.4	42.7
Despersonalización	n	189	158	164
	%	37.0	30.9	32.1
Realización Personal	n	81	121	309
	%	15.9	23.7	60.5

Fuente: Matriz de datos

GRÁFICO 2

PORCENTAJES DE ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 – II SEGÚN PUNTUACIÓN POR COMPONENTE DEL SÍNDROME DE BURNOUT



Fuente: Tabla 7

INTERPRETACIÓN: En la Tabla 8 y Gráfico 2 observamos que en el componente de Cansancio Emocional, el 42.7% de los estudiantes presentaron una puntuación considerada Baja, en el componente de Despersonalización, el 37% de los estudiantes

presento una puntuación Alta, y por último el 60.5% de los estudiantes presentaron una Baja realización personal.

DISCUSIÓN: Al respecto podemos observar que en la investigación de Carmen López, Ángela Zegarra y Víctor Cuba se registró que el componente de Cansancio emocional la mayor puntuación fue Baja con un 61.4%, lo que concuerda con nuestra investigación, en el componente de Despersonalización se identificó que el 50% presentó un Bajo nivel de Despersonalización, lo que no concuerda con la investigación, dado que se registro una Alta despersonalización en los estudiantes, por último en el componente de Realización personal en la investigación antes mencionada se encontró una Alta realización personal con un 72.7%, lo que no concuerda con nuestra investigación.

TABLA 9

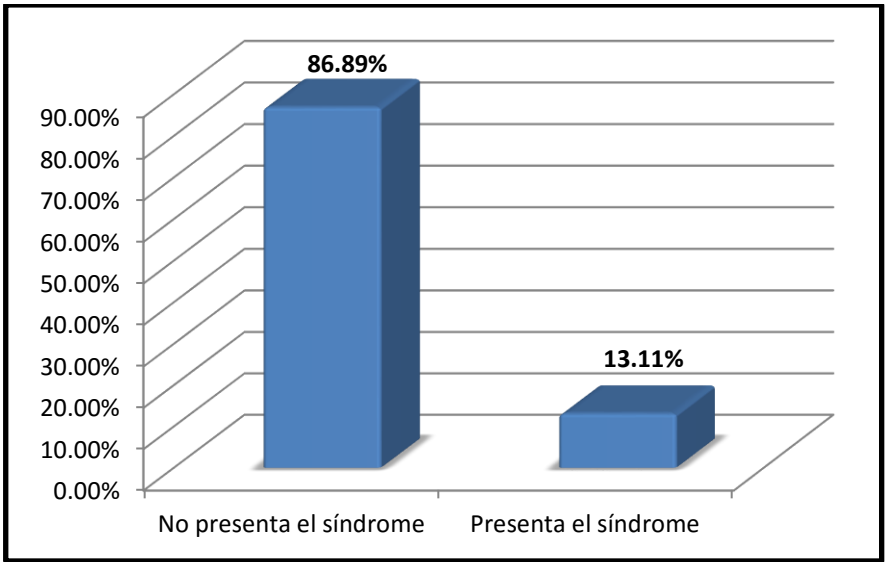
PORCENTAJES Y NÚMERO DE ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 – II SEGÚN PRESENCIA DEL SÍNDROME DE BURNOUT

SÍNDROME DE BURNOUT	n	%
No presenta el síndrome	444	86.89
Presenta el síndrome	67	13.11
Total	511	100

FUENTE: Matriz de datos

GRÁFICO 3

PORCENTAJE DE ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013-II SEGÚN PRESENCIA DEL SÍNDROME DE BURNOUT



Fuente: Tabla 8

INTERPRETACIÓN: En la Tabla 9 y el Gráfico 3 podemos observar que el 13.11% de los estudiantes presentaron el Síndrome de Burnout.

DISCUSIÓN: En la investigación de Carmen López, Ángela Zegarra y Víctor Cuba se encontró que el 13.6% de la enfermeras sufrían del Síndrome de Burnout, resultado similar al de la presente investigación, en contradicción a nuestro resultado se encuentra el obtenido por Elena Fernández, quien encuentra que la presencia de Síndrome de Burnout, en los estudiantes universitarios es elevada y muy fuerte, la contradicción puede deberse a la muestra de estudio, dado que

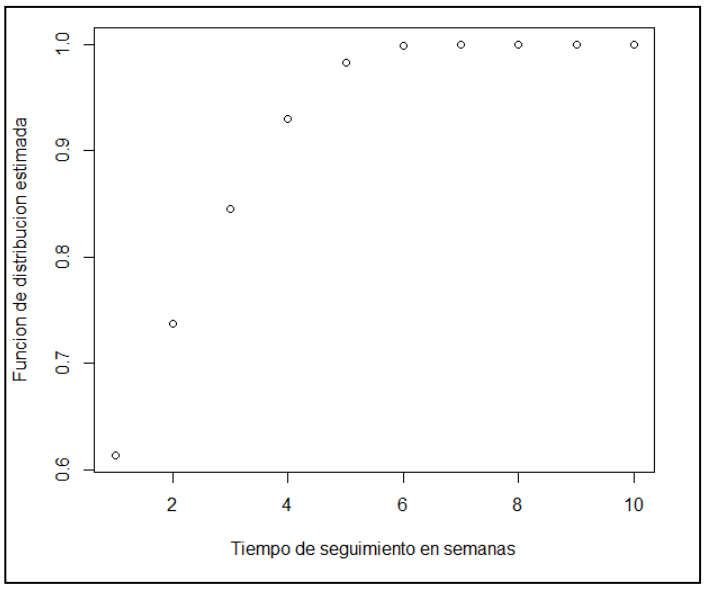
la investigación de Fernández, sólo consideró a dos escuelas profesionales, Enfermería y Topografía.

4.1.2 Estimación de modelos de regresión

La distribución de las probabilidades de la Presencia del Síndrome de Burnout en los estudiantes de la Universidad Nacional de Altiplano, se basan en la continuidad de las variables a través de la recta numérica, por lo que se realizó la estimación de la función de distribución para el Síndrome de Burnout (Anexo N° 09) valores necesarios para la estimación y prueba de censura de datos.

GRÁFICO 4

FUNCIÓN DE DENSIDAD DE LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 - II SEGÚN TIEMPO DE SEGUIMIENTO



FUENTE: Matriz de datos (Anexo Nº 09)

INTERPRETACIÓN: En el Gráfico 4 podemos observar que la función de distribución estimada para el Síndrome de Burnout, tiene un crecimiento marcado en las primeras semanas del seguimiento, y a partir de la quinta semana la probabilidad tiende a estar próxima a uno, por lo que podemos notar que esta distribución tiene cierta semejanza con el gráfico del clásico modelo de regresión logística.

A partir de estas observaciones podemos proceder a la prueba de la censura de datos.

TABLA 10

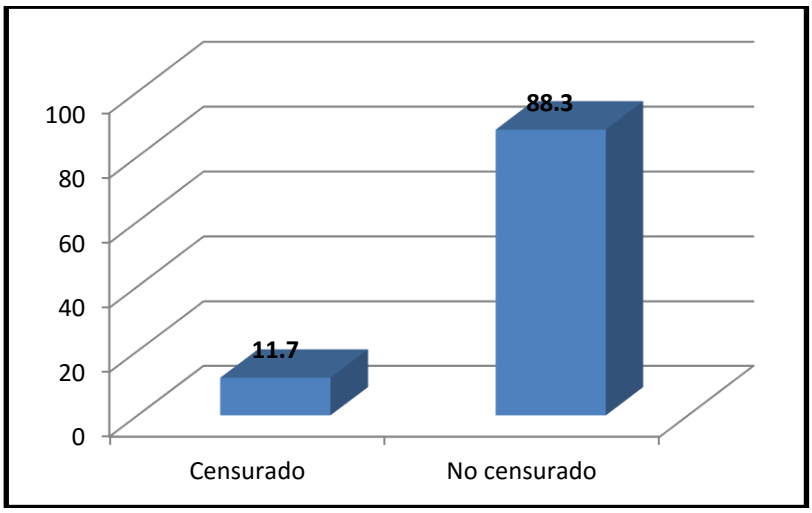
PROCENTAJE Y NÚMERO DE DATOS CENSURADOS Y NO CENSURADOS PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013-II

ESTADO	n	%
Censurado	60	11.7
No censurado	451	88.3

FUENTE: Matriz de datos

GRÁFICO 5

PORCENTAJE DE DATOS CENSURADOS Y NO CENSURADOS PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013-II

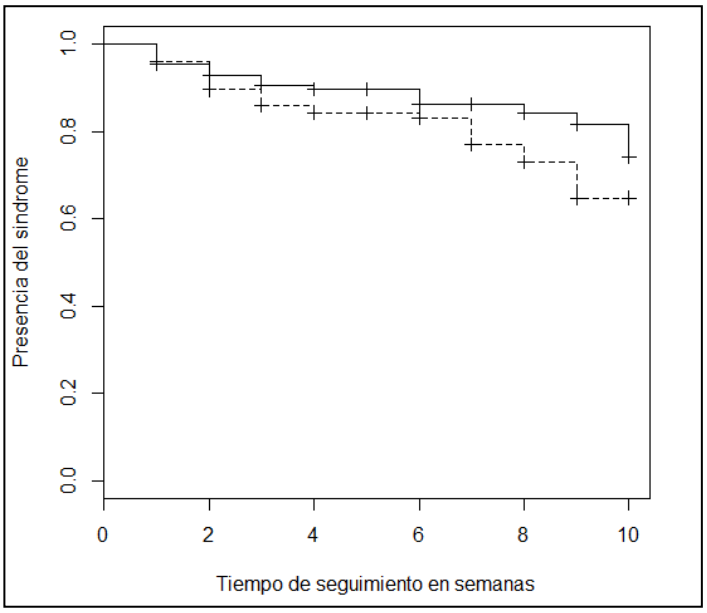


FUENTE: Tabla 9

INTERPRETACIÓN: En la Tabla 10 y Gráfico 5, se observa el número de datos censurados, los que representan el 11.7% de los datos, en la investigación de Castillo, López, Ramos y Fernández, especifica que el número de datos censurados no debe superar el 12% del total para garantizar la robustez de los modelos que se estimen a partir de estos datos, el valor obtenido para la presente investigación está por debajo de esta valor, lo que garantiza que el modelo con datos censurados, puede presentar estimaciones robustas.

GRÁFICO 6

FUNCIÓN DE DISTRIBUCIÓN DE LA CENSURA DE DATOS PARA LA ESTIMACIÓN DE LA PRESENCIA EL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013-II



Fuente: Matriz de datos

INTERPRETACIÓN: En el Gráfico 6 observamos una tendencia decreciente de la función de distribución de los datos censurados, podemos apreciar que en cada semana a partir del inicio de seguimiento hasta la décima semana hubo casos de Síndrome de Burnout, o también llamados fallos en el análisis de supervivencia, además la tendencia descendente de de la función de distribución de los datos censurados puede ser explicada por la disminución del número de elementos muestrales a través del tiempo de seguimiento.

TABLA 11

VALORES ESTIMADOS DEL MODELO DE REGRESIÓN LOGÍSTICA CON DATOS CENSURADOS PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013-II

VARIABLES	B	WALD	SIG.	EXP(B)
Sexo	,521	3,571	,059	1,683
Edad	-,256	,921	,337	,774
Tiempo de estudio	,603	14,342	,000	1,827
Lugar de procedencia	,436	9,490	,002	1,546
Gasto mensual	-1,295	38,468	,000	,274
Lugar de residencia	,111	,093	,760	1,117
Situación de convivencia	-,001	,000	,996	,999
Propiedad de la residencia	,364	4,729	,030	1,439

Fuente: Matriz de datos

INTERPRETACIÓN: En la Tabla 11 se observa las estimaciones para el modelo de regresión logística con datos censurados los factores más influyentes, fueron: Tiempo de estudio, Lugar de procedencia, Gasto mensual y Propiedad, todas estas variables presentaron una $p < 0.05$, lo que demuestra que son significantes y que tiene una influencia en la presencia del Síndrome de Burnout.

Se observó las variables más influyentes por lo que se procedió a recalcular el modelo para estas variables obteniéndose las siguientes estimaciones:

TABLA 12

VALORES ESTIMADOS DEL MODELO DE REGRESIÓN LOGÍSTICO CON DATOS CENSURADOS DE LAS VARIABLES MÁS INFLUYENTES PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013-II

VARIABLE	B	WALD	SIG.	EXP(B)
Tiempo de estudio	,559	13,043	,000	1,749
Lugar de procedencia	,445	9,882	,002	1,560
Gasto mensual	-1,288	38,960	,000	,276
Propiedad de la residencia	,424	7,467	,006	1,528

Fuente: Matriz de datos

INTERPRETACIÓN: En la Tabla 12 podemos observar que todas las variables tienen una significancia menor al 5%, lo que refleja que todas están relacionas con la presencia del Síndrome de Burnout, quedando el modelo matemático estimado de la siguiente manera:

$$p_i = \frac{1}{1 + \exp(-1.749x_{3i} - 1.560x_{4i} - 0.276x_{5i} - 1.528x_{8i})}$$

Donde:

x_{3i}: Es la medición de la variable Tiempo de estudio para el i–ésimo sujeto.

x_{4i}: Es la observación de la variable Lugar de procedencia para el i–ésimo sujeto.

x_{5i}: Es la observación de la variable Gasto mensual para el i–ésimo sujeto.

x_{8i}: Es la observación de la variable Situación de residencia para el i–ésimo sujeto.

TABLA 13

VALORES ESTIMADOS DEL MODELO DE REGRESIÓN LOGÍSTICA SIN DATOS CENSURADOS PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013-II

VARIABLES	B	WALD	SIG.	EXP(B)
Sexo	,815	1,135	,062	2,259
Edad	,041	,011	,915	1,042
Tiempo de estudio	1,196	23,370	,000	3,306
Lugar de procedencia	,727	16,716	,000	2,069
Gasto mensual	-1,776	47,811	,000	,169
Lugar de residencia	,319	,314	,576	1,375
Situación de convivencia	,187	,518	,472	1,206
Propiedad de la residencia	,722	5,429	,020	2,060

Fuente: Matriz de datos

INTERPRETACIÓN: En la Tabla 13 se observa las estimaciones para el modelo de regresión logística sin datos censurados, los factores más influyentes en la presencia del Síndrome de Burnout, fueron: Tiempo de estudio, Lugar de procedencia, Gasto mensual y Propiedad de la residencia, estos factores presentaron una $p < 0.05$ lo que representa que tienen una influencia significativa en la presencia del síndrome de Burnout.

TABLA 14

VALORES ESTIMADOS DEL MODELOS DE REGRESIÓN LOGÍSTICO CON DATOS NO CENSURADOS DE LAS VARIABLES MÁS INFLUYENTES PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013-II

VARIABLE	B	WALD	SIG.	EXP(B)
Tiempo de estudio	1,100	21,337	,000	3,005
Lugar de procedencia	,741	18,383	,000	2,099
Gasto mensual	-1,731	51,117	,000	,177
Propiedad de la residencia	,841	9,037	,003	2,318

Fuente: Matriz de datos

INTERPRETACIÓN: En la Tabla 14 podemos observar que todas las variables tienen una significancia menor al 5%, lo que refleja que todas están relacionas con la presencia del Síndrome de Burnout, quedando el modelo matemático estimado de la siguiente manera:

$$p_i = \frac{1}{1 + \exp(-3.005x_{3i} - 2.099x_{4i} - 0.177x_{5i} - 2.318x_{8i})}$$

Donde:

x_{3i}: Es la medición de la variable Tiempo de estudio para el i–ésimo sujeto.

x_{4i}: Es la observación de la variable Lugar de procedencia para el i–ésimo sujeto.

x_{5i}: Es la observación de la variable Gasto mensual para el i–ésimo sujeto.

x_{8i}: Es la observación de la variable Situación de residencia para el i–ésimo sujeto.

COMPARACIÓN DE LA INFLUENCIA DE LAS COVARIABLES EN AMBOS MODELOS

Como podemos observar en ambos modelos las variables Tiempo de estudio, Lugar de procedencia, Gasto mensual y la Propiedad de la residencia, fueron significativas para la presencia del síndrome de Burnout, lo que indica que existe una concordancia entre ambos modelos.

DISCUSIÓN: En la investigación de Carmen López, Ángela Zegarra y Víctor Cuba, se encontró que estas variables no tuvieron relación con la presencia del síndrome de Burnout, la discrepancia con los resultados es esta investigación, en la investigación de Elena Fernández, se encontró que la presencia del síndrome de Burnout en los estudiantes estaba relacionado con el Sexo, lo que no concuerda con la presente investigación, puesto que en ambos modelos esta variable demostró ser no significativa para la presencia del síndrome de Burnout, también notamos que existen discrepancias en cuanto a la Situación de convivencia y el tiempo de estudio de los estudiantes.

4.1.3 Comparación de los modelos de regresión

Se podreció a la elección de mejor parámetro de suavizado.

TABLA 15

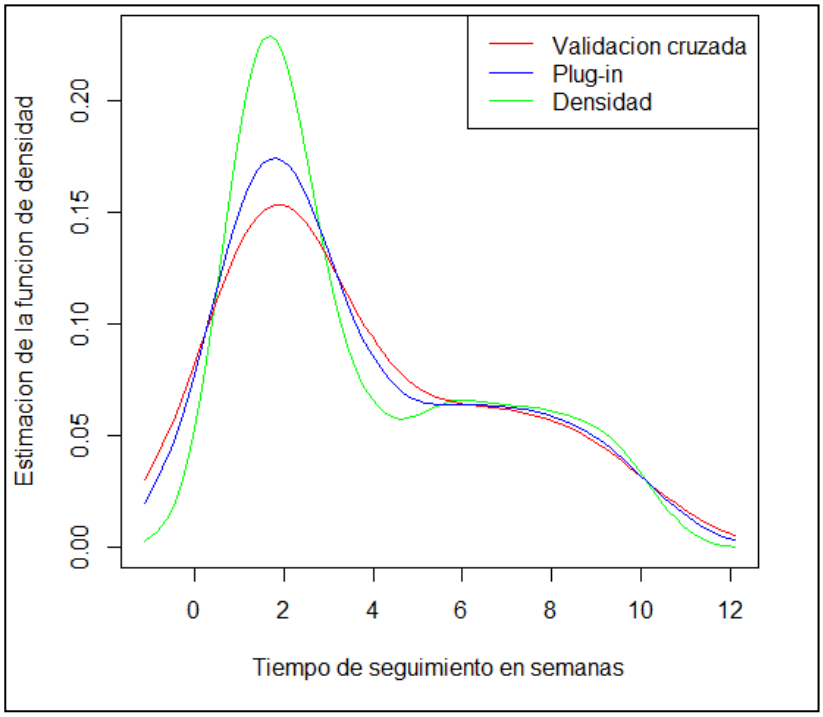
VALORES DE LOS PARÁMETROS DE SUAVIZACIÓN ESTIMADOS PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 - II

TIPO DE ESTIMACIÓN	VALOR ESTIMADO
Validación cruzada	1.49023
Plug- in (Ruppert, Sheather y Wand)	1.239042
Plug- in (Polansky y Baker) o densidad	0.8078978

Fuente: Matriz de datos

GRÁFICO 7

FUNCIÓN DE DENSIDAD CON DISTINTOS PARÁMETROS DE SUAVIZACIÓN ESTIMADOS PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 - II



Fuente: Matriz de datos(Anexo N°04 , Anexo N°05 , Anexo N°06)

INTERPRETACIÓN: En el Gráfico 7 podemos observar que no existen diferencias notorias entre el parametro de suavizado seleccionado por el método de Validacion cruzada y el método Plug in (Ruppert, Sheather y Wand), sin embargo ambos estimadores son diferentes del estimador Plug-in (Polansky y Baker) de densidad (Anexo N°04 , Anexo N°05 , Anexo N°06). Se observa a partir de la quinta semana una diferencia notoria en las funciones de densidad lo que se explica por la influencia de las covariables en la presencia del sindrome de Burnout.

TABLA 16

VALORES AMISE DE LOS PARÁMETROS DE SUAVIZACIÓN ESTIMADOS PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 - II

TIPO DE ESTIMACIÓN	VALOR DE AMISE
Validación cruzada	0.4537928
Plug- in (Ruppert, Sheather y Wand)	0.7446333
Plug- in (Polansky y Baker) o densidad	0.09762879

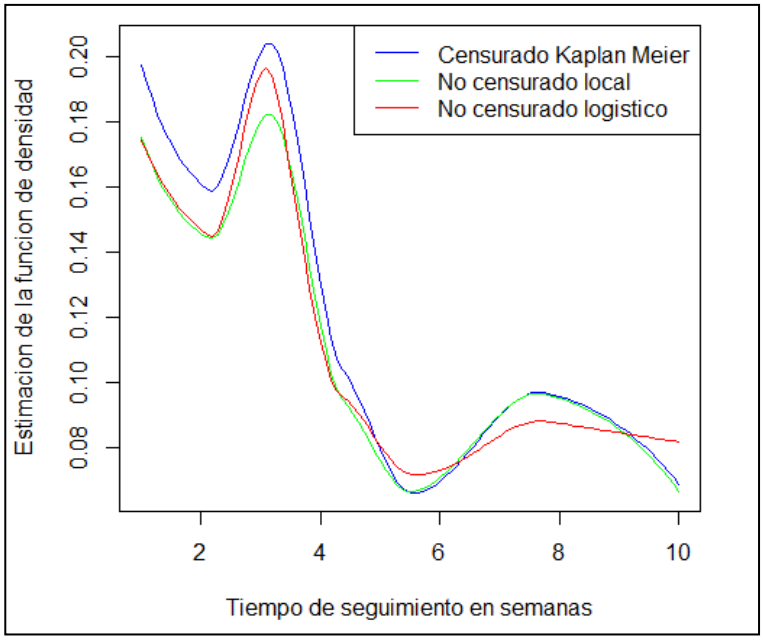
Fuente: Matriz de datos

INTERPRETACIÓN: Como se observa en la Tabla 16 el parámetro de suavizado Plug-in (Polansky y Baker) estima un menor error medio integrado (0.09762879) en comparacion con los otros estimadores, por lo tanto es el que se utilizó para la estimación de las funciones de densidad de los modelos de regresión local lineal, modelos de regresión logística sin datos

censurados y basado en el estimador Kaplan Meier ó modelo de regresión logística con datos censurados.

GRÁFICO 8

FUNCIÓN DE DENSIDAD DE LOS MODELOS DE REGRESIÓN PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 – II



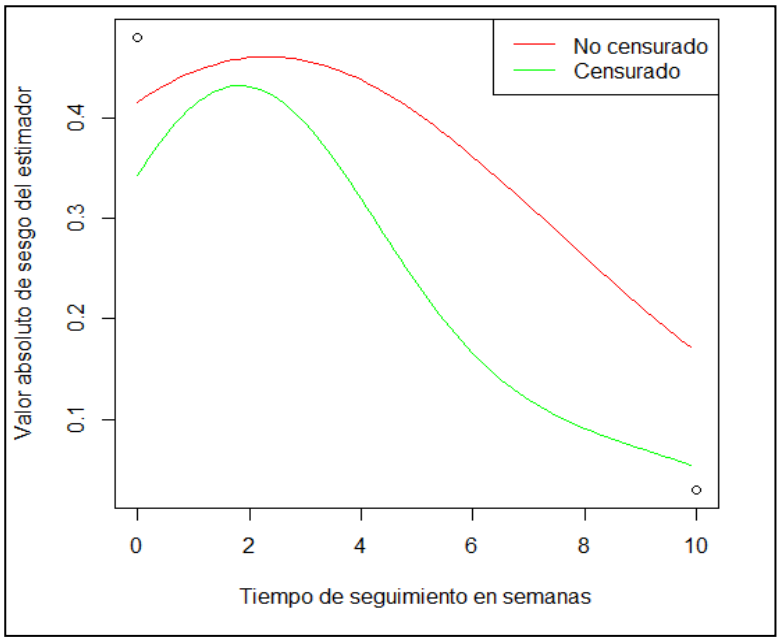
Fuente: Matriz de datos

INTERPRETACIÓN: Examinando el Gráfico 8 observamos que existe una semejanza entre el estimador de regresión local logistica y el estimador de regresión local hasta la semana 2 del estudio, sin embargo a partir de la semana 4 se observa que existen diferencias marcadas entre ambos estimadores. Evaluando las diferencias entre los modelos con datos censurados y no censurados, observamos que a partir de la quinta semana de seguimiento existe semejanza entre el estimador no censurado local y el estimador Kaplan Meier.

También podemos notar que la forma de las funciones denota cierta dependencia de las covariables.

GRÁFICO 9

VALOR ABSOLUTO DEL SESGO ASINTÓTICO ESTIMADO PARA LAS DENSIDADES DE LOS MODELOS DE REGRESIÓN LOGISTICA PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 - II



Fuente: Matriz de datos

INTERPRETACIÓN: En el Grafico 9 se representa el valor absoluto de las estimaciones de las expresiones asintóticas de los sesgos de ambos estimadores (Anexo N°07, Anexo N°08) de la probabilidad de sufrir el Síndrome de Burnout. Puesto que claramente se observa que el sesgo asintótico del estimador con datos censurados es mucho menor que el estimador con datos no censurados, está demostrado que este estimador representa un mejor ajuste para la presencia del síndrome de Burnout.

Además del sesgo asintótico, podemos obtener la pseudoverosimilitud y el error medio integrado para ambas funciones, estos estadísticos nos indicaron con mayor certeza cuál de ambos modelos de regresión es mejor.

TABLA 17

MEDIDAS DE PSUEDOVEROSILITUD Y ERROR MEDIO INTEGRADO PARA LA ESTIMACIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL DEL ALTIPLANO PUNO, 2013 - II

MEDIDA	MODELO CON DATOS NO CENSURADOS	MODELO CON DATOS CENSURADOS
Pseudoverosimilitud	1.949805	2.939877
Error medio integrado	0.0446937	0.0410464

Fuente: Matriz de datos

INTERPRETACIÓN: En la Tabla 17 se observa claramente que el modelo que incluye el factor de censura, presenta un menor error de estimación, además de tener una mayor pseudoverosimilitud con la distribución de los datos, lo que indica que el modelo que incluye los datos censurados es mejor que el modelo sin datos censurados, para la estimación de la probabilidad de la presencia del síndrome de Burnout de los estudiantes de la Universidad Nacional del Altiplano 2013 - II.

DISCUSIÓN: En la investigación de Amalia Jácome en la aplicación a los datos reales, concluye que el estimador de regresión logístico sin datos censurados estima un menor sesgo que el modelo de regresión logístico con datos censurados, lo que no concuerda con los resultados obtenidos en la presente

investigación, esta discrepancia puede ser aplicada por la influencia de las variables independientes. En la investigación de Castillo, López, Ramos y Fernández se concluye que la inclusión de los datos censurados en el modelos produce mejores estimaciones, resultados que son concordantes con los aquí hallados.

CONCLUSIONES

- El 13.11% de los estudiantes de la Universidad Nacional del Altiplano Puno 2013 – II sufren del Síndrome de Burnout.

- Los factores que mostraron influencia significativa ($p<0.05$) en la presencia del Síndrome de Burnout para los estudiantes de la Universidad Nacional del Altiplano fueron: Tiempo de estudio, Lugar de procedencia, Gasto mensual y Propiedad de la residencia.

- El modelo de regresión logística con datos censurados que tuvo mayor significancia ($p=0.00$) fue:

$$p_i = \frac{1}{1 + \exp(-1.749x_{3i} - 1.560x_{4i} - 0.276x_{5i} - 1.528x_{8i})}$$

- El modelo de regresión logística con datos no censurados que tuvo mayor significancia ($p=0.00$) fue:

$$p_i = \frac{1}{1 + \exp(-3.005x_{3i} - 2.099x_{4i} - 0.177x_{5i} - 2.318x_{8i})}$$

- El mejor parámetro de suavizado para la estimación de las funciones de densidad de ambos modelos tuvo un valor de 0.8078978 el que se estimó con la metodología Plug – in (Polansky y Baker) puesto que presentó un error de 0.09762879, valor que fue mucho menor en comparación con otros parámetros seleccionados.

- El modelo de regresión logística con datos censurados basado en el estimador Kaplan Meier obtuvo menor sesgo asintótico en comparación con el modelo de regresión logística con datos no censurados.

- El modelo de regresión logística con datos censurados basado en el estimador Kaplan Meier obtuvo un error medio integrado de 0.0410464 menor al 0.0446937 obtenido por el modelo de regresión logística con datos no censurados.

- El modelo de regresión logística con datos censurados basado en el estimador Kaplan Meier obtuvo una pseudoverosimilitud de 2.939877 mayor a la obtenida por el modelo de regresión logística con datos no censurados 1.949805.

- El mejor modelo para la estimación del síndrome de Burnout fue el modelo de regresión logística con datos censurados, cuya ecuación fue: $p_i = \frac{1}{1+\exp(-1.749x_{3i}-1.560x_{4i}-0.276x_{5i}-1.528x_{8i})}$, puesto que mostro un menor sesgo asintótico y error medio integrado, además de mayor pseudoverosimilitud con respecto al modelo de regresión logística con datos no censurados.

RECOMENDACIONES Y SUGERENCIAS

- Se recomienda tener en cuenta que la salud emocional y psicológica es un factor importante en la vida estudiantil, por lo que las autoridades de nuestra casa superior de estudios deberían tomar en cuenta este tipo de información para implementar medidas que contribuyan con la mejora de la salud emocional y psicológia de los estudiantes de la Universidad Nacional del Altiplano.

- Se recomienda poder realizar un estudio donde se mida los efectos de los factores sociales, teniendo en cuenta los distintos periodos de tiempo como la inicialización y finalización del semestre académico.

- Se recomienda a las personas interesadas en las estimaciones de tipo no paramétrico como funciones núcleo (Kernel), realizar aplicaciones del estimador polinomial local, o los estimadores de Nadayara Watson, dado que en esta investigación no se desarrolló por cuestiones de la distribución de los datos.

- En la presente investigación se realizó la recolección de datos con el test Maslach Burnout Inventory (Ver Anexo N°03), sin embargo existe muchos otros instrumentos de medición para este síndrome, por lo que se recomienda realizar investigación con distintos tipos de instrumentos tales como el Cuestionario de Sentido de Coherencia de Antonovsky, o el Cuestionario de Estrés Percibido (CEP) de Sanz-Carrillo.

BIBLIOGRAFÍA

ADERSEN, P. K., BORGAN, O., GILL, R. D., & KEIDING, N. (1993). *Statistical models based on counting process.* New York: Springer - Verlag .

BAE, J., & KIM, S. (2003). The uniform law of large numbers for the Kaplan- Meier integral process. *Bulletin of the Aurtralian Mathematical Society* , 459-465.

BLEMA, M. J., & TOBLAS, A. (2001). *Aplicación de los modelos de regresión tobit en la modelizacion de variables epidemiológicas censuradas.* Madrid.

BLUM, J. R., & SUSARLA, V. (1980). Maximal deviation theory of density and failure rate function estimates based on censored data. *Journal of Multivariate Analysis* , 213-222.

BOCHNER, S. (1955). *Harmonic analysis and the theory of probability.* California.

BOTELLA, P., ROCAMORA, M. A., MARTÍNEZ, B., & ALACRÉU, M. (2001). *Introducción al Análisis de Supervivencia.* Barcelona: Universidad Cardenal Herrera.

BOUKICHOU ABDELKADER, N. (2010). *Regresion no paramétrica en R.* Granada: Universidad de Granada.

BOWMAN, A. W. (1985). A comparative study of some kernel-based nonparametric density estimators. *J. Statistic and Comput. Simul.* , 313-327.

BOWMAN, A. W. (1984). An alternative method of cross validation for the smoothing of density estimates. *Biometrika* , 353-360.

CACERES, R. A. (2007). *Estadística aplicada a las ciencias de la salud.* España: DIAZ DE SANTOS.

CANNON, W. B. (1929). *BODILY CHANGES IN PAIN, HUNGER, FEAR AN RAGE .* New York: Appleton.

CARREÑO, A. (2006). *Análisis de superviviencia.* Barcelona: Diaz de santos.

CASTILLO, E., LÓPEZ, M., FERNÁNDEZ, A., & RAMOS, A. (2006). *Influencia del número de datos censurados en la evaluación del campo S-N.* UNIVERSIDAD DE CANTABRIA.

CAYUELA, L. (2010). Análisis multivariante. Andalucía: Centro Andaluz de Medio Ambiente.

CLEVELAND, W. S. (1979). Robust Loacally Weighted Regression and Smmothing Scatterplots. In *Journal of the American Statistical Association* (pp. 829-836).

CSÖRGO, S., & HORVÁTH, L. (1983). The rate of the strong uniform consistency for the product-limit estimator. In S. CSÖRGO, & L. HORVÁTH, *Zeetschrift für Wahrscheinlichketstheorie und verwandte* (pp. 411-426). Gebiete.

DIKTA, G. (1998). On semiparametric random censorship models. *Journal of Statistics ans Planning Ingerence* , 253-279.

DIKTA, G. (2000). The strong law under semiparametric random censorship models . *Journal of Statisrtics and Planning Inference* , 1-10.

DIKTA, G. (2001). Weak representation of the cumulative hazard function under semiparametric random censorship models. In G. DIKTA, *Statistics* (pp. 395-409).

DOMINGUEZ, E., & ALDANA, D. (2001). Regresión Logistica: un ejemplo de su uso en la endocrinologia. *Revista Cubana Endocrinal* , 58-64.

EFRON, B. (1967). The two-sample problem with censored data. *Proceddings Fifth Berkeley Symposium Math. Statistics and Probability* , 831-853.

FARBER, B. (1983). *Stress and burnout in the human service professions.* Nueva York: Pergamon Press.

FERNÁNDEZ MARTÍNEZ, M. E. (2009). *Estrés percibido, estrategias de afrontamiento y sentido de coherencia en estudiantes de enfermería: su asociación con la salud psicológica y estabilidad emocional.* LEÓN: UINVERSIDAD DE LEÓN.

FLEMING, T. R., & HARRINGTON, D. P. (1991). *Countind process and survival analysis.* New York: Wiley.

FÖLDES, A., & RETJÖ, L. (1980). Strong uniform cinsistency for nonparametric estimation based on censored data. In *The Annals of Statistics* (pp. 122-129).

GAN, Y., SHANG, J., & ZHANG, Y. (2007). Coping flexibility and locus of control as predictors of burnout among chinese college students. In Y. GAN, J. SHANG, & Y. ZHANG, *Social Behavior ans Personalituy* (pp. 1087-1098).

JÁCOME, M. A. (2005). *Estimación presuavizada de las funciones de densidad y distribución con datos censurados.* Universidade da Coruña.

JOHANSEN, S. (1978). The product limit estimator as maximun likelihood estimator. *Scandinavian Journal of Statistics* , 195-199.

JOMES, M. C., MARRON, J. S., & SHEATHER, S. J. (1996). Abrief sruvey of bandwith selection for density estimation. JASA.

KAPLAN, E. L., & MEIER, P. (1958). Nonparametric estimation from incomplete observations. In *Journal of the American Satatistical Association* (pp. 457-481).

KIVIMÄKI, M., VAHTERA, J., ELOVAINIO, M., LILLRANK, B., & KEVIN, M. (2002). *Death or illness of a family memeber, violence, interpersonal conflict, and financial difficulties as predictors of sickness absence: Longitudinal cohort study on psychological behavioral links.* Psychosomatic Medicine.

LAZARUS, R. S., & FOLKMAN, S. (1987). Transactional theory and research on emotions and coping. *European Journal f Personality* , 147-169.

LO, S. H., & SINGH, K. (1986). The product-limit estimator and the bootstrap ; some asymptotic representations. In S. H. LO, & K. SINGH, *Probability Theory ans Related Fields* (pp. 455-465).

LÓPEZ, C., ZEGARRA, Á., & CUBA, V. (2006). Factores Asociados al Síndrome de Burnout en Enfermeras del Emergencia del Hospital Nacional Guillermo Almenara Irigoyen. *Revista de Ciencias de la Salud* , 53-61.

MARTÍNEZ FALERO, J. E., AYUGA TÉLLEZ, E., & GONZÁLES GARCÍA, C. (1992). *Onteción del mejor ajuste según el tipo de datos.* Madrid: E.T.S.I.

MASLACH, C., & JACKSON, S. E. (1981). *Maslasch Burnout Inventory. Manuel.* Palo Alto, California: Consulting Psychologist Press.

MEDINA, E. (2003). *Modelos de Elección Discreta.* Barcelona: Diaz de santos.

MIÑARRO, A. (1998). *Estimación no paramétrica de la función de densidad.* Barcelona: Diaz de santos.

Nadaraya, E. (1989). *Nonparametric Estimation of Probability Densities and Regression Curve.* Dordrecht: Kluwer Academic Publishers.

PARRA MURCIEGO, J. M. (2011). *Estimación de un modelo aditivo no paramétrico.* Granada.

PHADIA, E. G., & SHAO, P. Y. (1999). Exact moments of the product limit estimator. In *Statistics and Probability Letters* (pp. 277-286).

REBASA, P. (2005). Conceptos básicos del analisis de supervivencia. *Análisis de supervivencia* , 222-230.

ROSENBLATT, M. (1956). Remarks on some non-parametrics estimates of a density function. *Annals Math. Statist.* , 832-837.

RUDEMO, M. (1982). Empirical choise of histograms and kernel density estimators. *Scandinavian Journal of Statistics* , 65-78.

SALANOVA, M., GRAU, R., & MARTÍNEZ, I. (2005). *Jop demands and coping behaviour: the moderating role of professional self- efficacy.* Psicothema.

SALCEDO, C. M. *Estimación de la ocurrencia de incidencias de declaraciones de pólizas de importación.* Lima: Universidad Mayor de San Marcos.

SELYE, H. (1973). *The evolution of th stress concept.* American Science.

SHORACK, G., & WELLNER, J. A. (1986). *Empirical processes with applications to statistics.* New York: Wiley.

SILVERMAN, B. W. (1986). *Density estimation for statistics and data analysis.* Londres: Chapman and Hall.

STUTE, W. (1995). The central limit theorem under random censorship. In W. STUTE, *The Annals of Statistics* (pp. 422-439).

STUTE, W., & WANG, J. L. (1993). Astorng law under random censorship. In W. STUTE, & J. L. WANG, *The Annals of Statistics* (pp. 1591-1607).

TUSELL, T. (2012). *Análisis multivariante.* Barcelona: Diaz de santos.

WEBGRAFÍA

http://www.psicologia-online.com/ebooks/riesgos/capitulo4_5.shtml

Hora: 19:33 Fecha: 28/01/2013

http://www.psicologia-online.com/ebooks/riesgos/capitulo4_4.shtml

Hora 19:38 Fecha: 28/01/2013

http://www.castalia.org.uy/docs/libros/DrograsyEtica/FreddyDaSilva.pdf

Hora: 07:30 Fecha: 29/01/2013

http://www.hrc.es/bioest/Reglog_1.html

Hora 15:15 Fecha: 09/05/2013

http://www.seh-lelha.org/rlogis1.htm

Hora 15:07 Fecha: 10/05/2013

http://lbe.uab.es/vm/salud/apuntes/logistica.pdf

Hora: 10:35 Fecha: 13/05/2013

https://www.editorialdigitaltec.com/materialadicional/P019_Arroyo_Analisismultivariante_cap1.pdf

Hora: 6:30 Fecha: 18/05/2013

http://www.acmcb.es/files/425-3397-DOCUMENT/Sancho-9-14Maig12.pdf

Hora: 6:52 Fecha: 18/05/2013

http://www.slideshare.net/tecnomexico/analisis-multivariable

Hora: 7:14 Fecha: 18/05/2013

http://www.uam.es/personal_pdi/ciencias/jspinill/documentos/METODOS/Metodos304011_Intro_y_GLM.pdf

Hora: 11:15 Fecha 18/10/2013

http://webmelisa.es/docs/xornadar/R_nunha_Unidade_de_Investigaci%C3%B3n_Biom%C3%A9dica.pdf

Hora: 19:05 Fecha 18/10/2013

http://imarrero.webs.ull.es/sctm05/modulo2lp/10/psaavedra.pdf

Hora 07:00 Fecha 19/10/2013

http://masteres.ugr.es/moea/pages/tfm1011/estimacionnoparametricadeunmodeloaditivo

Hora: 6:41 Fecha 8/11/2013

http://www.uam.es/personal_pdi/economicas/eva/pdf/logit.pdf

Hora: 6:45 Fecha: 8/11/2013

http://www.hrc.es/bioest/Reglog_10.html

Hora: 7:05 Fecha: 8/11/2013

http://catarina.udlap.mx/u_dl_a/tales/documentos/lec/pineda_r_g/capitulo5.pdf

Hora: 7:16 Fecha: 8/11/2013

http://www.geogra.uah.es/firemap/pdf/VilardelHoyo_Granada.pdf

Hora 7:20 Fecha: 8/11/2013

http://www.uv.es/armero/temes_msuavitzats/tema2.pdf

Hora 11:13 Fecha 21/11/2013

http://www.uv.es/~armero/temes_msuavitzats/tema1.pdf

Hora 11:23 Fecha 21/11/2013

http://e-archivo.uc3m.es/handle/10016/15537?locale-attribute=en

Hora 11:29 Fecha: 21/11/2013

http://www.uv.es/conesa/CursoR/material/handout-sesion3.pdf

Hora 11:35 Fecha: 21/11/2013

ANEXOS

ANEXO N° 01

UNA/P: ESTUDIANTES MATRICULADOS EN PRIMER SEMESTRE 2013 POR SEXO, SEGÚN FACULTAD Y ESCUELA PROFESIONAL

Nª	FACULTAD	ESCUELA PROFESIONAL	H	M	TOTAL
1	CIENCIAS AGRARIAS	Ingeniería Agronómica	245	102	**347**
		Ingeniería Agroindustrial	142	153	**295**
		Ingeniería Topográfico y Agrimensura	377	53	**430**
2	MEDICINA VETERINARIA Y ZOOTECNIA	Medicina Veterinaria y Zootecnia	537	181	**718**
3	INGENERÍA ECONÓMICA	Ingeniería Económica	433	297	**730**
4	CIENCIAS CONTABLES Y ADMINISTRATIVAS	Ciencias Contables	453	466	**919**
		Administración	348	333	**681**
5	ENFERMERÍA	Enfermería	80	459	**539**
6	TRABAJO SOCIAL	Trabajo Social	86	460	**546**
7	INGENERÍA DE MINAS	Ingeniería de Minas	678	11	**689**
8	CIENCIAS SOCIALES	Sociología	280	223	**503**
		Turismo	216	285	**501**
		Antropología	204	136	**340**
		Ciencias de la Comunicación Social	311	268	**579**
		Arte	305	81	**386**
9	BIOLOGÍA	Biología	231	195	**426**
10	CIENCIAS DE LA EDUCACIÓN	Educación Secundaria	478	343	**821**
		Educación Física	174	29	**203**
		Educación Primaria	96	230	**326**
		Educación Inicial	27	270	**297**
11	INGENIERÍA ESTADÍSTICA E INFORMÁTICA	Ingeniería Estadística e Informática	239	85	**324**
12	INGENIERÍA GEOLÓGICA Y METALÚRGICA	Ingeniería Geológica	506	76	**582**
		Ingeniería Metalúrgica	297	56	**353**
13	CIENCIAS JURÍDICAS Y POLÍTICAS	Derecho	418	234	**652**
14	INGENIERÍA QUÍMICA	Ingeniería Química	156	124	**280**
15	CIENCIAS DE LA SALUD	Nutrición Humana	83	247	**330**
		Odontología	277	222	**499**
16	INGENIERÍA AGRICOLA	Ingeniería Agrícola	384	95	**479**
17	INGENIERÍA CIVIL Y ARQUITECTURA	Ingeniería civil	754	50	**804**
		Arquitectura y Urbanismo	405	190	**595**
		Ciencias Físico - Matemáticas	146	21	**167**
18	MEDICINA HUMANA	Medicina Humana	270	150	**420**
19	INGENIERIA MECÁNIA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS	Ingeniera Mecánica Eléctrica	626	5	**631**
		Ingeniería Electrónica	415	13	**428**
		Ingeniería de Sistemas	455	89	**544**
TOTAL			**11132**	**6232**	**17364**
Porcentaje			**64,11**	**35,89**	**100**

Fuente: Oficina de Tecnología Informática – Información al 24 de abril de 2013
Sistematización y Elaboración: UNA- OGPD- Oficina de Estadística

ANEXO N° 02

UNIVERSIDAD NACIONAL DEL ALTIPLANO
FACULTAD DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA
Escuela Profesional de Ingeniería Estadística e Informática

La presente encuesta se realiza con un fin investigativo, por lo que se ruega pueda brindar la información de la manera más apropiada, agradeciendo enormemente el tiempo y la importancia que se presta para poder llevar a cabo la recopilación de la información necesaria para desarrollar esta investigación. Se le pide que llene o marque según corresponda los siguientes datos:

DATOS GENERALES:
Llene la información en los espacios en blanco

Escuela Profesional	
Semestre que cursa	

DATOS PERSONALES:
Llene la información en los espacios en blanco o marque la opción que más se adapte a su realidad.

NOMBRES Y APELLIDOS:

Sexo

a) Femenino
b) Masculino

¿Cuál es su edad?

_____

¿Cuál es su estado civil?

a) Soltero
b) Casado
c) Conviviente
d) Divorciado
e) Otros

¿Cuál es su lugar de procedencia?

_____

¿Vive actualmente en Puno?
a) Si
b) No

¿Con quién vive en la actualidad?
a) Vivo solo
b) Vivo con mis padres y/o hermanos
c) Vivo con otros familiares

ANEXO N° 03

TEST DEL SÍNDROME DE BURNOUT

Los siguientes 22 ítems son referentes a sus sentimientos con respecto a los estudios que cursa en la universidad, lea cuidadosamente la declaración e indique con un aspa o una x en la casilla que mejor represente cuan seguido se presentan estos sentimientos de acuerdo a la siguiente escala:

0	1	2	3	4	5	6
Nunca	Pocas veces al año	Una vez al mes	Más de una vez al mes	Una vez a la semana	Más de una vez a la semana	Todos los días

AFIRMACIONES	0	1	2	3	4	5	6
1. Me siento emocionalmente agotado en la universidad.							
2. Me siento cansado al final de cada día en la universidad.							
3. Me siento fatigado cuando me levanto por la mañana y tengo que ir a la universidad.							
4. Comprendo fácilmente como se sienten mis compañeros y docentes.							
5. Creo que trato a algunos compañeros y docentes como si fueran objetos y no personas.							
6. Estudiar todo el día con mucha gente es un gran esfuerzo para mí.							
7. Trato muy eficazmente los trabajos que me dejan los docentes en la universidad.							
8. Me siento agotado por el estudio.							
9. Creo que influyo positivamente con mi estudio en la vida de las personas que me rodean.							
10. Me he vuelto más insensible con la gente desde que estudio esta profesión.							
11. Me preocupa el hecho de que el estudio me endurezca emocionalmente.							
12. Me siento muy activo.							
13. Me siento frustrado con mis estudios.							
14. Creo que estoy estudiando demasiado.							
15. Realmente no me preocupa lo que le ocurre a mis compañeros y docentes.							
16. Estudiar con mis compañeros y/o tratar con los docentes me estresa.							
17. Puedo crear fácilmente una atmósfera relajada con mis compañeros y docentes.							
18. Me siento estimulado después de estudiar con mis compañeros y docentes.							
19. He conseguido muchas cosas útiles estudiando mi carrera.							
20. Me siento acabado.							
21. En la universidad trato los problemas emocionales con mucha calma.							
22. Siento que mis compañeros y docentes me culpan por alguno de sus problemas.							
TOTAL							

ANEXO N° 04

VALORES DE LA FUNCIÓN DE DENSIDAD PARA EL PARÁMETRO DE SUAVIZADO OBTENIDO POR EL MÉTODO DE VALIDACIÓN CRUZADA

0.030573594	0.035382541	0.040652549	0.046372181	0.052518663
0.059057142	0.065940297	0.073108377	0.080489696	0.088001629
0.095552081	0.103041438	0.110364926	0.117415314	0.124085869
0.130273438	0.135881539	0.140823332	0.145024329	0.148424733
0.150981297	0.152668617	0.153479814	0.153426563	0.152538487
0.150861948	0.148458298	0.145401667	0.141776415	0.137674345
0.133191802	0.128426788	0.123476196	0.118433264	0.113385336
0.108411985	0.103583548	0.098960082	0.094590748	0.090513608
0.086755786	0.083333963	0.080255143	0.077517628	0.075112155
0.073023114	0.071229806	0.069707694	0.068429580	0.067366702
0.066489706	0.065769474	0.065177807	0.064687952	0.064274981
0.063916031	0.063590412	0.063279616	0.062967234	0.062638801
0.062281608	0.061884482	0.061437563	0.060932093	0.060360234
0.059714923	0.058989773	0.058179019	0.057277517	0.056280796
0.055185141	0.053987718	0.052686721	0.051281528	0.049772860
0.048162927	0.046455542	0.044656209	0.042772153	0.040812304
0.038787215	0.036708932	0.034590789	0.032447164	0.030293183
0.028144389	0.026016392	0.023924509	0.021883410	0.019906786
0.018007051	0.016195088	0.014480054	0.012869236	0.011367981
0.009979684	0.008705836	0.007546127	0.006498599	0.005559828

ANEXO N° 05

VALORES DE LA FUNCIÓN DE DENSIDAD, PARA EL PARÁMETRO DE SUAVIZADO OBTENIDO POR EL MÉTODO DE PLUG IN

0.019897675	0.024371801	0.029540594	0.035434490	0.042066730
0.049429754	0.057492089	0.066196017	0.075456286	0.085160073
0.095168373	0.105318878	0.115430318	0.125308140	0.134751294
0.143559786	0.151542604	0.158525579	0.164358697	0.168922455
0.172132851	0.173944727	0.174353271	0.173393610	0.171138555
0.167694679	0.163197003	0.157802641	0.151683831	0.145020734
0.137994436	0.130780479	0.123543232	0.116431282	0.109573993
0.103079230	0.097032247	0.091495596	0.086509942	0.082095593
0.078254581	0.074973097	0.072224154	0.069970309	0.068166358
0.066761899	0.065703697	0.064937815	0.064411448	0.064074458
0.063880580	0.063788302	0.063761406	0.063769210	0.063786512
0.063793286	0.063774171	0.063717809	0.063616083	0.063463317
0.063255488	0.062989489	0.062662489	0.062271409	0.061812525
0.061281210	0.060671812	0.059977652	0.059191154	0.058304067
0.057307793	0.056193789	0.054954047	0.053581606	0.052071107
0.050419335	0.048625735	0.046692861	0.044626721	0.042436987
0.040137045	0.037743863	0.035277666	0.032761437	0.030220243
0.027680442	0.025168802	0.022711579	0.020333619	0.018057533
0.015902988	0.013886153	0.012019332	0.010310788	0.008764753
0.007381629	0.006158323	0.005088710	0.004164181	0.003374221

ANEXO N° 06

VALORES DE LA FUNCIÓN DE DENSIDAD, PARA EL PARÁMETRO DE SUAVIZADO OBTENIDO POR EL MÉTODO DE PLUG- IN (POLANSKY Y BAKER).

0.0029531191	0.0046020763	0.0069861964	0.0103331968	0.0148951186
0.0209312194	0.0286834758	0.0383461910	0.0500328149	0.0637444345
0.0793450705	0.0965485934	0.1149206716	0.1338969017	0.1528156668
0.1709619534	0.1876169211	0.2021077350	0.2138529492	0.2224001257
0.2274537978	0.2288928632	0.2267768647	0.2213405953	0.2129765245
0.2022051267	0.1896345069	0.1759125442	0.1616765707	0.1475066830
0.1338886531	0.1211909190	0.1096576284	0.0994168178	0.0905002777
0.0828700638	0.0764462370	0.0711311598	0.0668272007	0.0634465346
0.0609134417	0.0591607890	0.0581230959	0.0577287310	0.0578934803
0.0585171162	0.0594838364	0.0606666697	0.0619352434	0.0631657371
0.0642514512	0.0651122398	0.0657011379	0.0660068841	0.0660516884
0.0658844355	0.0655703828	0.0651791253	0.0647729508	0.0643976168
0.0640770391	0.0638125390	0.0635863510	0.0633682633	0.0631237250
0.0628215898	0.0624398733	0.0619684160	0.0614080549	0.0607666699
0.0600531511	0.0592707848	0.0584116767	0.0574535910	0.0563600210
0.0550835766	0.0535720492	0.0517759916	0.0496564389	0.0471915052
0.0443809443	0.0412482131	0.0378399874	0.0342233826	0.0304812987
0.0267063960	0.0229942622	0.0194363867	0.0161136084	0.0130907017
0.0104126766	0.0081031668	0.0061649911	0.0045826548	0.0033262882
0.0023563512	0.0016284101	0.0010973933	0.0007209237	
0.0004615503				

ANEXO N° 07

VALORES ESTIMADOS DEL SESGO ASINTÓTICO PARA LA FUNCIÓN DE DENSIDAD DEL MODELO DE REGRESIÓN LOCAL LOGÍSTICO SIN DATOS CENSURADOS

0.4154242	0.4190930	0.4226285	0.4260271	0.4292854	0.4324000
0.4353676	0.4381852	0.4408500	0.4433592	0.4457103	0.4479010
0.4499290	0.4517926	0.4534897	0.4550189	0.4563788	0.4575682
0.4585860	0.4594316	0.4601043	0.4606038	0.4609299	0.4610826
0.4610622	0.4608690	0.4605038	0.4599674	0.4592609	0.4583853
0.4573423	0.4561333	0.4547602	0.4532250	0.4515297	0.4496767
0.4476685	0.4455077	0.4431972	0.4407398	0.4381386	0.4353968
0.4325179	0.4295052	0.4263623	0.4230930	0.4197010	0.4161901
0.4125645	0.4088280	0.4049848	0.4010392	0.3969953	0.3928574
0.3886299	0.3843170	0.3799233	0.3754531	0.3709107	0.3663007
0.3616275	0.3568954	0.3521088	0.3472722	0.3423898	0.3374660
0.3325051	0.3275112	0.3224886	0.3174413	0.3123734	0.3072889
0.3021917	0.2970857	0.2919746	0.2868621	0.2817518	0.2766472
0.2715517	0.2664687	0.2614013	0.2563527	0.2513259	0.2463239
0.2413494	0.2364052	0.2314939	0.2266180	0.2217798	0.2169817
0.2122258	0.2075143	0.2028491	0.1982320	0.1936648	0.1891492
0.1846867	0.1802788	0.1759268	0.1716319		

ANEXO N° 08

VALORES ESTIMADOS DEL SESGO ASINTÓTICO PARA LA FUNCIÓN DE DENSIDAD DEL MODELO DE REGRESIÓN LOCAL LOGÍSTICO SIN DATOS CENSURADOS

0.34246637	0.35109970	0.35943940	0.36745506	0.37511708
0.38239679	0.38926673	0.39570077	0.40167430	0.40716445
0.41215020	0.41661252	0.42053457	0.42390178	0.42670195
0.42892537	0.43056487	0.43161589	0.43207653	0.43194752
0.43123228	0.42993683	0.42806982	0.42564240	0.42266821
0.41916321	0.41514564	0.41063581	0.40565604	0.40023043
0.39438473	0.38814614	0.38154315	0.37460529	0.36736300
0.35984737	0.35208994	0.34412253	0.33597699	0.32768506
0.31927810	0.31078697	0.30224182	0.29367194	0.28510556
0.27656977	0.26809032	0.25969158	0.25139635	0.24322584
0.23519957	0.22733530	0.21964901	0.21215487	0.20486522
0.19779058	0.19093964	0.18431935	0.17793489	0.17178979
0.16588594	0.16022370	0.15480197	0.14961830	0.14466895
0.13994902	0.13545256	0.13117266	0.12710159	0.12323087
0.11955143	0.11605371	0.11272773	0.10956324	0.10654982
0.10367694	0.10093411	0.09831092	0.09579713	0.09338277
0.09105819	0.08881410	0.08664165	0.08453250	0.08247878
0.08047321	0.07850906	0.07658020	0.07468110	0.07280684
0.07095309	0.06911614	0.06729284	0.06548064	0.06367750
0.06188192	0.06009290	0.05830988	0.05653275	0.05476177

ANEXO N° 09

VALORES ESTIMADOS DE LA FUNCIÓN DE DISTRIBUCIÓN PARA EL SÍNDROME DE BURNOUT

```
[1] 1.0000000 1.0000000 1.0000000 1.0000000 0.9299160 1.0000000 1.0000000
 [8] 1.0000000 1.0000000 0.9299160 1.0000000 0.7369103 0.9831275 0.7369103
[15] 0.8451226 0.7369103 1.0000000 0.7369103 0.7369103 0.6134423 0.7369103
[22] 0.6134423 1.0000000 1.0000000 0.8451226 0.9831275 0.6134423 1.0000000
[29] 1.0000000 0.9986289 1.0000000 0.9831275 1.0000000 1.0000000 0.9986289
[36] 0.7369103 0.7369103 0.7369103 0.6134423 0.9831275 0.6134423 0.6134423
[43] 0.9986289 0.7369103 0.7369103 0.7369103 1.0000000 0.7369103 0.7369103
[50] 0.6134423 1.0000000 1.0000000 1.0000000 0.6134423 1.0000000 1.0000000
[57] 1.0000000 0.8451226 0.9831275 0.9986289 0.9986289 1.0000000 0.9986289
[64] 0.7369103 0.9831275 0.7369103 0.7369103 0.8451226 0.6134423 0.9986289
[71] 0.9299160 0.8451226 0.6134423 0.6134423 0.6134423 1.0000000 0.6134423
[78] 0.6134423 0.7369103 1.0000000 0.7369103 0.8451226 0.9299160 0.7369103
[85] 0.8451226 0.6134423 0.7369103 0.9986289 0.7369103 0.7369103 0.8451226
[92] 0.7369103 0.6134423 0.9831275 0.6134423 1.0000000 0.9299160 0.8451226
[99] 0.8451226 0.7369103 0.7369103 0.6134423 0.9986289 1.0000000 1.0000000
[106] 1.0000000 0.7369103 0.6134423 0.8451226 0.7369103 0.7369103 0.6134423
[113] 0.9831275 0.7369103 0.7369103 0.7369103 0.7369103 0.6134423 0.7369103
[120] 1.0000000 1.0000000 1.0000000 1.0000000 0.9986289 1.0000000 1.0000000
[127] 1.0000000 1.0000000 1.0000000 0.7369103 0.8451226 1.0000000 0.8451226
[134] 0.9831275 0.7369103 0.6134423 0.9831275 0.6134423 1.0000000 1.0000000
[141] 1.0000000 0.9831275 1.0000000 0.6134423 0.9986289 1.0000000 1.0000000
[148] 1.0000000 1.0000000 1.0000000 0.9986289 1.0000000 0.6134423 0.6134423
[155] 0.7369103 0.9986289 0.7369103 0.8451226 0.7369103 1.0000000 0.9986289
[162] 0.9986289 1.0000000 0.6134423 0.7369103 0.8451226 0.7369103 0.6134423
[169] 0.7369103 0.7369103 0.7369103 0.7369103 0.6134423 0.9831275 0.9986289
[176] 0.6134423 0.6134423 0.6134423 0.9986289 0.6134423 0.6134423 0.6134423
[183] 0.6134423 0.6134423 0.6134423 0.7369103 0.7369103 0.7369103 0.7369103
[190] 0.7369103 0.7369103 0.7369103 0.7369103 0.9986289 1.0000000 1.0000000
[197] 0.9986289 1.0000000 0.9986289 0.9986289 0.6134423 1.0000000 0.9299160
[204] 0.9986289 1.0000000 1.0000000 0.6134423 1.0000000 0.9831275 0.6134423
[211] 0.7369103 0.9299160 0.7369103 0.9831275 0.9299160 0.7369103 0.6134423
[218] 0.9986289 0.7369103 1.0000000 1.0000000 1.0000000 1.0000000 0.6134423
[225] 0.9831275 0.7369103 0.6134423 0.7369103 0.8451226 0.7369103 1.0000000
[232] 0.8451226 0.7369103 0.9299160 1.0000000 0.6134423 0.9986289 0.9986289
[239] 0.6134423 0.6134423 0.6134423 1.0000000 1.0000000 0.9986289 0.6134423
[246] 1.0000000 0.6134423 0.6134423 0.6134423 0.6134423 0.6134423 0.7369103
[253] 0.7369103 1.0000000 0.7369103 0.7369103 0.7369103 0.7369103 0.6134423
[260] 0.7369103 0.6134423 0.9831275 1.0000000 0.6134423 0.6134423 0.6134423
[267] 1.0000000 0.9299160 0.9299160 0.8451226 0.8451226 0.7369103 0.7369103
```

CONTINUACIÓN DEL ANEXO N° 09

[274] 0.7369103 0.9986289 0.7369103 0.8451226 0.6134423 0.6134423 0.6134423
[281] 0.6134423 0.7369103 0.9299160 0.7369103 0.6134423 1.0000000 0.6134423
[288] 1.0000000 0.9299160 1.0000000 1.0000000 1.0000000 0.9831275 0.6134423
[295] 1.0000000 0.8451226 0.9299160 0.7369103 0.7369103 0.6134423 0.7369103
[302] 0.7369103 1.0000000 0.8451226 0.9831275 0.8451226 0.7369103 0.6134423
[309] 0.7369103 0.7369103 0.9299160 0.7369103 0.6134423 0.9986289 0.6134423
[316] 0.7369103 1.0000000 0.9299160 1.0000000 0.9299160 1.0000000 0.7369103
[323] 0.9986289 1.0000000 1.0000000 1.0000000 0.9299160 0.6134423 0.7369103
[330] 0.6134423 1.0000000 0.7369103 0.7369103 0.7369103 0.6134423 0.6134423
[337] 0.6134423 0.6134423 0.9299160 0.7369103 0.7369103 0.6134423 0.9299160
[344] 0.6134423 0.7369103 0.9299160 0.8451226 0.8451226 0.6134423 0.9831275
[351] 0.6134423 0.6134423 0.7369103 0.7369103 0.6134423 0.6134423 0.6134423
[358] 0.6134423 0.6134423 0.6134423 1.0000000 0.9299160 0.7369103 1.0000000
[365] 0.7369103 0.7369103 0.7369103 0.9986289 1.0000000 0.6134423 0.7369103
[372] 0.7369103 0.6134423 1.0000000 1.0000000 0.9986289 0.6134423 0.6134423
[379] 0.6134423 0.9831275 0.7369103 0.6134423 1.0000000 0.6134423 0.6134423
[386] 0.7369103 0.7369103 0.9299160 0.6134423 0.9831275 0.7369103 1.0000000
[393] 0.7369103 0.6134423 0.6134423 0.6134423 1.0000000 0.7369103 0.7369103
[400] 0.8451226 0.6134423 0.7369103 1.0000000 0.7369103 0.7369103 0.7369103
[407] 0.9299160 0.7369103 0.8451226 0.8451226 0.8451226 0.9831275 0.8451226
[414] 0.9986289 0.9831275 0.7369103 0.7369103 0.8451226 0.8451226 1.0000000
[421] 0.9986289 0.7369103 1.0000000 0.9299160 0.9831275 1.0000000 0.8451226
[428] 0.8451226 0.8451226 0.9299160 1.0000000 0.7369103 0.6134423 0.7369103
[435] 0.7369103 1.0000000 0.7369103 0.9831275 0.9831275 1.0000000 0.7369103
[442] 0.7369103 0.8451226 0.7369103 0.6134423 1.0000000 0.9986289 0.6134423
[449] 0.6134423 0.6134423 0.6134423 0.6134423 0.6134423 0.6134423 0.6134423
[456] 0.8451226 1.0000000 0.7369103 0.7369103 0.7369103 0.7369103 0.7369103
[463] 0.7369103 0.8451226 0.7369103 0.9986289 1.0000000 0.7369103 0.8451226
[470] 0.7369103 0.9299160 0.8451226 0.7369103 0.7369103 0.7369103 0.9986289
[477] 0.7369103 0.8451226 0.7369103 0.6134423 1.0000000 0.7369103 0.7369103
[484] 0.9986289 0.6134423 0.6134423 0.6134423 0.6134423 0.6134423 0.6134423
[491] 0.6134423 1.0000000 1.0000000 0.9986289 1.0000000 0.6134423 1.0000000
[498] 0.7369103 0.6134423 0.7369103 0.6134423 1.0000000 0.9986289 0.7369103
[505] 0.7369103 1.0000000 0.6134423 0.7369103 0.7369103 0.6134423 1.0000000

ANEXO N° 10

DISTRIBUCIÓN DE LA PRESENCIA DEL SÍNDROME DE BURNOUT SEGÚN ESCUELA PROFESIONAL

Escuela Profesional	No presenta en síndrome		Presenta el síndrome		Total
	N	%	n	%	n
Ingeniería Agronómica	9	90.0%	1	10.0%	10
Ingeniería Agroindustrial	8	88.9%	1	11.1%	9
Ingeniería Topográfica y Agrimensura	8	61.5%	5	38.5%	13
Medicina Veterinaria y Zootecnia	16	76.2%	5	23.8%	21
Ingeniería Económica	19	90.5%	2	9.5%	21
Ciencias Contables	23	85.2%	4	14.8%	27
Administración	19	95.0%	1	5.0%	20
Enfermería	11	68.8%	5	31.3%	16
Trabajo Social	15	93.8%	1	6.3%	16
Ingeniería de Minas	20	100%	0	0%	20
Sociología	15	1005	0	0%	15
Turismo	12	80.0%	3	20.0%	15
Antropología	10	100%	0	0%	10
Ciencias de la Comunicación Social	14	82.4%	3	17.6%	17
Arte	11	100%	0	0%	11
Biología	10	76.9%	3	23.1%	13
Educación Secundaria	19	79.2%	5	20.8%	24
Educación Física	4	66.7%	2	33.3%	6
Educación Primaria	9	90.0%	1	10.0%	10
Educación Inicial	8	88.9%	1	11.1%	9
Ingeniería Estadística e Informática	9	100%	0	0%	9
Ingeniería Geológica	16	94.1%	1	5.9%	17
Ingeniería Metalúrgica	9	90.0%	1	10.0%	10
Derecho	14	73.7%	5	26.3%	19
Nutrición Humana	7	87.5%	1	12.5%	8
Odontología	10	100.0%	0	0%	10
Ingeniería Agrícola	13	86.7%	2	13.3%	15
Ingería Civil	14	100%	0	0%	14
Ingeniería Civil	21	87.5%	3	12.5%	24
Arquitectura y Urbanismo	17	100%	0	0%	17
Ciencias Físico Matemáticas	4	80.0%	1	20.0%	5
Medicina Humana	7	58.3%	5	41.7%	12
Ingeniería Mecánica Eléctrica	16	84.2%	3	15.8%	19
Ingeniería Electrónica	12	92.3%	1	7.7%	13
Ingeniería de Sistemas	15	93.8%	1	6.3%	16

Fuente: Matriz de datos

ANEXO N° 11

CÁLCULO DE LA MUESTRA

Calculando la proporción general

$$P_g = \sum_{h=1}^{L} W_h P_h$$

$$P_g = \frac{347(0.8) + 295(0.75) + 430(0.72) + \cdots + 544(0.78)}{17364}$$

$$P_g = 0.724330707$$

Cálculo de error muestral

$$E = e * P_g$$

$$E = 0.05 * 0.724330707$$

$$E = 0.036216535$$

Cálculo del tamaño de muestra

$$n_0 = \frac{\frac{1.96^2[347(0.8 * 0.2) + 295(0.75 * 0.25) + 430(0.72 * 0.28) + \cdots + 544(0.788 * 0.212)]}{17364}}{(0.05 * 0.03622)^2}$$

$$n_0 = 510.4157387 \approx 511$$

Condición para la corrección de la muestra inicial

Como $\frac{511}{17364} = 0.02939506 < 0.05$, entonces no se corrige la muestra inicial