Divulgaciones Matemáticas Vol. 17 No. 2 (2016), pp. 43–56

Generalización del estimador de Kaplan-Meier para tiempos de vida difusos

Generalization of the Kaplan-Meier estimator for fuzzy lifetimes

José A. González C. (jgonzalez@upla.cl)
Laboratorio de investigación Lab[e]saM, Departamento de Matemática y Estadística
Universidad de Playa Ancha, Chile.

Resumen

Esta propuesta entrega una generalización del estimador de Kaplan-Meier, en la cual los tiempos de vida son considerados números difusos. Esta propuesta se sitúa en un contexto mucho más real para el proceso de medición, considerando la imprecisión propia de la naturaleza humana. Es necesario para ello la definición de algunos conceptos como frecuencia relativa y clases difusas. Se presentan algunos resultados asintóticos y finalmente una aplicación y comparación con la metodología convencional de estimación.

Palabras y frases clave: conjunto difusos, número difuso, función de pertenencia, Kaplan-Meier, asintótico.

Abstract

This proposal provides a generalization of the Kaplan-Meier estimator, where lifetimes are considered fuzzy numbers. This proposal is consistent with a much more realistic context to the measurement process considering the imprecision of human nature. To achieve our goal, concepts such as relative frequency and fuzzy classes are given. Moreover, some asymptotic results are presented, together with an application and comparison with the conventional estimation method.

Key words and phrases: fuzzy set, fuzzy number, membership function, Kaplan-Meier, asymptotic.

1 Introducción

La importancia del estimador producto límite o Kaplan-Meier es indiscutible, sin embargo los valores numéricos usados para representar las mediciones son idealizaciones de informaciones imprecisas. Esos tipos de imprecisiones están relacionados con el proceso de medición, ya sea en el registro y cuantificación del tiempo de vida. La certeza o verdad, la duda o incerteza, han estado en el centro de la epistemología del conocimiento. En la filosofía griega, Platón localiza la certeza en el "mundo de las ideas", en oposición al "mundo sensible", en la cual tendríamos solo aproximaciones de valores del mundo de las ideas.

Recibido 31/03/2016. Revisado 28/03/2017. Aceptado 19/07/2017.
MSC (2010): Primary 62N86 ; Secondary 20M20.
Autor de correspondencia: José A. González C.

La principal motivación de la teoría difusa es la construcción de una estructura cuantitativa formal capaz de capturar la imprecisión del conocimiento humano, esto es como ese conocimiento es formulado en un lenguaje natural. La teoría de conjuntos difusos es la conexión entre los modelos matemáticos tradicionales y la representación mental que es generalmente imprecisa [8].

En aplicaciones de ingenería, el tiempo de vida de un sistema puede ser dificil de medir, debido a la complejidad del sistema [13]. En esta situación no considerar la imprecisión puede llevar a resultados irreales [18]. Por tanto, los métodos para estimar la función no paramétrica de sobrevivencia usando el método de Kaplan-Meier debe ser adaptada a la situación de tiempos de vida difusos, con la finalidad de obtener resultados realistas. La teoría de conjuntos difusos es una importante herramienta matemática para lidiar con problemas de incerteza, imprecisiones o verdades parciales, permitiendo dar tratabilidad a problemas del mundo real, muchas veces con soluciones a bajo costo. Así la generalización del estimador de Kaplan-Meier considera los tiempos de vida como números difusos, sin embargo para el desarrollo del proceso de estimación bajo este concepto es necesario de la definición de algunos elementos, tales como frecuencia relativa y orden de los tiempos de vida.

2 Preliminares

Es importante indicar que son usados algunos conceptos y terminologías de la teoría de conjuntos difusos basados principalmente en los trabajos de [21], [8] y [15].

Definición 2.1. Conjunto difuso. Sea $\Omega \subseteq \mathbb{R}^k$ un subconjunto no vacío del espacio Euclidiano k-dimensional. Un conjunto difuso $\widetilde{A}$ es un conjunto de pares ordenados $\widetilde{A} = \left\{ \left(\omega, \mu_{\widetilde{A}}(\omega)\right) : \omega \in \Omega \right\}$, donde $\mu_{\widetilde{A}} : \Omega \to [0,1]$ es llamada función de pertenencia para el conjunto difuso $\widetilde{A}$. Adicionalmente, el conjunto difuso vacío $\widetilde{\varnothing}$ es caracterizado por $\mu_{\widetilde{\varnothing}}(\omega) = 0$ para todo $\omega \in \Omega \subseteq \mathbb{R}^k$.

Observemos que la teoría de conjutos difusos extiende la teoría de conjuntos convencionales, relajando el concepto de pertenencia de los elementos en sus respectivos conjuntos. Por un lado, en la teoría de los conjuntos convencionales se considera que $\omega \in A$ (pertenencia uno) o $\omega \notin A$ (pertenencia cero), es decir es una operación binaria. Por otro lado, la teoría de los conjuntos difusos considera un grado de pertenencia que varía en el intervalo $[0,1]$, es decir, ω es un elemento de A con un cierto grado y ese mismo elemento ω es también un elemento de A^c con otro grado. La teoría de probabilidad es contruída sobre la teoría de conjuntos usuales y provee un número en el intervalo $[0,1]$ para describir el grado de certeza que $\omega \in A$. La principal diferencia entre la teoría de probabilidades y la teoría de conjuntos difusos se encuentra en la defincón de un conjunto: la primera considera los conjunto tradicionales el otro considera los conjuntos difusos. Como veremos en esta sección, las propiedades de los conjuntos difusos son muy diferentes de las tradicionales. La aplicabilidad de los conjuntos difusos es enorme en la modelación linguistica[22], análisis de imagen, diseño de dispositivos electrónicos [9], etc.

De la Definición 2.1, podemos representar un conjunto común usando una notación difusa. Note que si $\Omega = \mathbb{R}^k$, cualquier subconjunto usual $B \subseteq \mathbb{R}^k$ es representado por definición por $\mu_{\widetilde{B}}(\omega) = 1$ para todo $\omega \in B$ y $\mu_{\widetilde{B}}(\omega) = 0$ para todo $\omega \notin B$. Por ejemplo, sea $\Omega = \mathbb{R}$ y $B = (a,b)$ un intervalo en la recta real con $a < b$. Entonces, B puede ser escrito en términos de un conjunto difuso como $\widetilde{B} = \{(\omega, 1) : \ \forall\, \omega \in B\} \cup \{(\omega, 0) : \ \forall\, \omega \notin B\}$.

Es importante resaltar que funciones de pertenencia y de densidad de probabilidad son intrinsecamente diferentes. Por ejemplo, si $\pi(\omega)$ es una función de densidad con $\pi(\omega) \geq 0$ para todo $\omega \in \Omega$ y $\int_\Omega \pi(\omega) d\omega = 1$, podemos obtener una función de pertenencia, definiendo

$\mu_{\widetilde{A}}(\omega) = C^{-1}\pi(\omega)$, desde que $C = \sup_{\omega\in\Omega} \pi(\omega) < \infty$. Sin embargo, el inverso no es necesariamente verdadero, debido a que una función de pertenencia no necesita ser integrable sobre Ω.

Para las funciones de densidad de probabilidad, es común definir un soporte para caracterizar el conjunto de todos los puntos con una densidad positiva. Para una función de pertenencia tenemos la misma definición para representar el conjunto de todos los puntos con pertenencia positiva en el conjunto difuso. La Definición 2.2 formaliza este concepto.

Definición 2.2. El soporte de un conjunto difuso $\widetilde{A}$ es definido como
$supp(\widetilde{A}) = \left\{\omega \in \Omega :\ \mu_{\widetilde{A}}(\omega) > 0\right\}$.

Note que un elemento ω tiene pertenencia plena en su respectivo conjunto difuso, cuando su pertenencia a él es uno. En este contexto el elemento ω contiene completamente todos los recursos exigidos por el conjunto difuso. La Definición 2.3 formaliza el conjunto de todos los puntos con pertenencia plena, esto es, todos los puntos donde sus funciones de pertenencia son iguales a 1.

Definición 2.3. El núcleo de un conjunto difuso $\widetilde{A}$ es definido como Nucleo$(\widetilde{A}) = \{\omega \in \Omega : \mu_{\widetilde{A}}(\omega) = 1\}$.

Cuando el núcleo tiene por lo menos un elemento tenemos un conjunto difuso normal, (ver Definición 2.4).

Definición 2.4. Un conjunto difuso $\widetilde{A}$ es llamado normal si su núcleo es no vacío. En otras palabras, existe por lo menos un punto $\omega \in \mathbb{R}^k$ con $\mu_{\widetilde{A}}(\omega) = 1$.

Sea $\widetilde{A}$ un conjunto difuso normal, luego mientras más próximo $\mu_{\widetilde{A}}(\omega_0)$ está de uno, más creemos que ω_0 se encuentra en core$(\widetilde{A})$ y mientras más próximo $\mu_{\widetilde{A}}(\omega_0)$ está de cero, más creemos que ω_0 no está en core$(\widetilde{A})$. Esto es, el grado de pertenencia de un elemento puede ser también visto como una medida de incerteza [22].

Sean $\widetilde{A}$ y $\widetilde{B}$ dos conjuntos difusos con función de pertenencia $\mu_{\widetilde{A}}(\omega)$ y $\mu_{\widetilde{B}}(\omega)$, respectivamente, de acuerdo con [21] [19], si $\Omega \subseteq \mathbb{R}^k$, entonces las operaciones comunes son definidas de la siguiente forma:

1. $\widetilde{A} \subseteq \widetilde{B} \Longleftrightarrow \mu_{\widetilde{A}}(\omega) \leq \mu_{\widetilde{B}}(\omega)$, para todo $\omega \in \Omega$.
2. $\widetilde{A} \equiv \widetilde{B} \Longleftrightarrow \mu_{\widetilde{A}}(\omega) = \mu_{\widetilde{B}}(\omega)$, para todo $\omega \in \Omega$.
3. $\widetilde{A}^c$ es el complemento de $\widetilde{A} \Longleftrightarrow \mu_{\widetilde{A}^c}(\omega) = 1 - \mu_{\widetilde{A}}(\omega)$, para todo $\omega \in \Omega$.
4. $\widetilde{C} = \widetilde{A} \cup \widetilde{B} \Longleftrightarrow \mu_{\widetilde{C}}(\omega) = \text{máx}\{\mu_{\widetilde{A}}(\omega), \mu_{\widetilde{B}}(\omega)\}$, para todo $\omega \in \Omega$.
5. $\widetilde{D} = \widetilde{A} \cap \widetilde{B} \Longleftrightarrow \mu_{\widetilde{D}}(\omega) = \text{mín}\{\mu_{\widetilde{A}}(\omega), \mu_{\widetilde{B}}(\omega)\}$, para todo $\omega \in \Omega$.

Basado en las definiciones anteriores, si consideramos $\widetilde{\Omega} = \{(\omega, 1);\ \omega \in \Omega \subseteq \mathbb{R}^k\}$ como el conjunto difuso universal, entonces para cualquier conjunto difuso $\widetilde{A}$ con función de pertenencia $\mu_{\widetilde{A}}(\omega)$ para todo $\omega \in \Omega$, tenemos que $\widetilde{A} \subseteq \widetilde{\Omega}$. Además si existe $\omega_0 \in \Omega$ tal que $\text{máx}\{\mu_{\widetilde{A}}(\omega_0), \mu_{\widetilde{A}^c}(\omega_0)\} < 1$ tenemos que $\widetilde{A} \cup \widetilde{A}^c \neq \widetilde{\Omega}$. Por otro lado, si existe $\omega_0 \in \Omega$ tal que $\text{mín}\{\mu_{\widetilde{A}}(\omega_0), \mu_{\widetilde{A}^c}(\omega)\} > 0$, tenemos que $\widetilde{A} \cap \widetilde{A}^c \neq \widetilde{\varnothing}$. Observemos que estas propiedades no atienden las leyes del tercero excluido y de contradicción de la teoría de los conjuntos clásicos [10, 19].

El concepto de conjunto difuso es muy amplio y difícil de lidiar sin algunas especificaciones adicionales. En este contexto la próxima definición permite especificar un número difuso, que es una extención natural de los números reales. Sin embargo, esta última definición depende de la convexidad en el contexto difuso, que se define a continuación.

Definición 2.5. [21] Un conjunto $\widetilde{A}$ es convexo, si y solo si

$$\mu_{\widetilde{A}}(\lambda\omega_1 + (1-\lambda)\omega_2) \geq \min\{\mu_{\widetilde{A}}(\omega_1), \mu_{\widetilde{A}}(\omega_2)\}$$

para todo $\omega_1, \omega_2 \in \Omega$ y $\lambda \in [0,1]$.

Observemos que el concepto de convexidad desde la perspectiva difusa difiere de la definición clásica de convexidad en análisis funcional. Un intervalo difuso $\widetilde{A}$ es un conjunto difuso que satisface la condición de convexidad y normalidad, de modo que el intervalo es construido por todos los elementos con función de pertenencia 1. Un intervalo difuso $\widetilde{A}$ es un número difuso cuando la cardinalidad del $Nucleo(\widetilde{A})$ es igual que 1 [15]. Los números e intervalos difusos son útiles para representar imprecisiones y medidas de intervalos, respectivamente. Estos conceptos tienen múltiples aplicaciones, por ejemplo, en inteligencia artificial, procesamiento de imágenes, reconocimiento de voz, ciencias biológicas y médicas, investigación de operaciones, análisis de decisión y procesamiento de información, economía, geografía, psicología, linguística, etc. Más aplicaciones pueden ser encontradas en [8] y [11].

En [8] fueron definidas las clases de funciones de pertenencia LR (left and right) definidas sobre $\Omega = \mathbb{R}$, es decir, la clase de funciones de pertenencia que pueden ser enteramente caracterizadas por tres parámetros, (m, α, β), y dos funciones L y R. La próxima definición está relacionada con el concepto de números difuso tipo LR.

Definición 2.6. El número difuso $\widetilde{A}$ es dicho tipo LR si existen dos funciones decrecientes $L, R : [0, +\infty) \to [0,1]$ con $L(0) = R(0) = 1$, $\lim_{\omega \to +\infty} L(\omega) = \lim_{\omega \to +\infty} R(\omega) = 0$ y números reales positivos $m \geq 0$, $\alpha > 0$, $\beta > 0$ tal que

$$\mu_{\widetilde{A}}(\omega) = \begin{cases} L\left(\dfrac{m-\omega}{\alpha}\right), & \text{para } \omega \leq m, \\ R\left(\dfrac{\omega - m}{\beta}\right), & \text{para } \omega \geq m, \end{cases}$$

donde m es llamado el centro de $\widetilde{A}$, α y β son llamados propagaciones izquierda y derecha respectivamente.

Si $\alpha = \beta$, $\widetilde{A}$ es llamado número difuso simétrico. Es importante resaltar que, para una función simétrica, se tiene la igualdad $L\left(\frac{m-\omega}{\alpha}\right) = R\left(\frac{\omega - m}{\beta}\right)$ para todo $\omega \in \mathbb{R}$. Si L y R son segmentos que comienzan en los puntos $(a_l, 0)$ y $(a_r, 0)$, respectivamente, y finalizan en $(a_m, 1)$, entonces se dice que $\widetilde{A}$ es un número difuso triangular.

3 Método

Cuando es asumida la naturaleza difusa de los tiempos de vida se refiere a la Definición 2.6. Por razones prácticas no tendría sentido tener soportes de amplitud infinita, ya que las expansiones son conocidas.

3.1 El estimador de Kaplan-Meier

El análisis de sobrevivencia y particularmente la metodología de estimación de Kaplan-Meier, es una importante línea de investigación, comenzando con un fuerte desarrollo aplicado, después desarrollos teóricos formales orientados al estudio de la normalidad asintótica del estimador [4], Extensiones dimensionales [6, 7, 17] y generalizaciones [1]. [2] estudia las propiedades de acotado y desigualdades del estimador de Kaplan-Meier, asociados con el estudio de varianza [16].

Actualmente el estimador de Kaplan-Meier es utilizado fuertemente en salud humana, permitiendo presentar mejores resultados, controles e identificación de los puntos de impacto significativo [20, 14], es actualmente una línea establecida de investigación.

La expresión general del estimador de Kaplan-Meier es presentada considerando los siguientes preliminares:

1. $t_1 < t_2 < ... < t_k$, son los tiempos diferentes ordenados de acuerdo con las fallas.
2. d_j el número de fallas en $t_j, j = 1, ..., k$, y
3. n_j el número de individuos en riesgo en el tiempo t_j, esto es, individuos que no fallaron y no fueron censurados hasta el momento inmediatamente antes de t_j.

Luego el estimador de Kaplan-Meier es definido como:

$$\hat{S}(t) = \prod_{j\,:\,t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j\,:\,t_j<t} \left(1 - \frac{d_j}{n_j}\right).$$

El estimador de Kaplan-Meier y otros estimadores tienen variaciones que deben ser descritas en términos de estimaciones intervalares. La expresión para la varianza asintótica del estimador de Kaplan-Meier, es dada por:

$$\hat{Var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j\,:\,t_j<t} \frac{d_j}{n_j(n_j - d_j)},$$

esta expresión es conocida como formula de Greenwood y es basada en las propiedades del estimador de máxima verosimilitud. Para t fijo, $\hat{S}(t)$ tiene distribución normal asintótica, luego un intervalo aproximado de $100(1-\alpha)\,\%$ de confianza para $S(t)$ es dado por:

$$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\hat{Var}(\hat{S}(t))}.$$

Esos elementos serán usados para comparar los resultados. Así para definir el estimador de Kaplan-Meier para tiempos de vida difusos, es necesario de dos desarrollos teóricos, la definción de un orden, el cual no será aborado en este artículo y la definición del concepto de frecuencia relativa en números difusos.

3.2 Frecuencia relativa difusa

El histograma es uno de los primeros y uno de los más comunes métodos de estimación de densidad. Es importante tener en mente que el histograma es una técnica de suavisamiento usado para estimar la densidad desconocida y por tanto merece alguna consideración. Se intenta reunir

los datos por contéo de cuántos datos o puntos pertenecen a un pequeño intervalo difuso de amplitud h. Este tipo de intervalos es llamado bin. Sin perdida de generalidad, se considera un bin centrado en 0, denominado intervalo difuso $\widetilde{H} = \{(x,1) : \frac{-h}{2} \leq x < \frac{h}{2}\}$.

La estimación natural de esta probabilidad es la frecuencia relativa de las observaciones en este intervalo de tiempo, es decir, se cuenta el número de observaciones en el intervalo y dividimos por el número total de observaciones, luego una generalización difusa debería ser similar.

Por tanto es necesario de algunos elementos base para la caracterización de un histograma en una estructura difusa.

Sea $\mathcal{M}$ una muestra observada y n el tamaño muestral, tal que $\mathcal{M} = \{\left(x, \mu_{\widetilde{A}}(x)\right) : \left(x, \mu_{\widetilde{A}}(x)\right) \in \widetilde{A}_j,\ \textit{para algun } j \in \{1,...,n\}\}$. También sea $\text{máx}(\mathcal{M}) = (\text{máx}(x), \text{mín}(\mu_{\widetilde{A}}(\text{máx}(x))))$ y $\text{mín}(\mathcal{M}) = (\text{mín}(x), \text{mín}(\mu_{\widetilde{A}}(\text{mín}(x))))$ el máximo y mínimo de la muestra, respectivamente.

Sea K el número de clases o intervalos de amplitud a_i, donde $i \in \{1,...,k\}$. L_{I_1} es el límite inferior de la primera clase, $L_{I_1} = \text{mín}(x) - \epsilon_1$, donde $\epsilon_1 > 0$ y L_{S_k} es el límite superior de la última clase, $L_{S_k} = \text{máx}(x) + \epsilon_2$, donde $\epsilon_2 > 0$. ϵ_1 y ϵ_2 dependen de la forma como obtenemos K, por ejemplo Sturges, $\sqrt{n}$, etc.

Sea C_i la h-ésima clase caracterizada por $C_i = \{x : (x, \mu_{\widetilde{A}}(x)) \in \mathcal{M}\ y\ L_{S_{i-1}} \leq x < L_{I_{i+1}}\}$ y $C_i \cap C_j = \varnothing,\ \forall i \neq j$.

Definición 3.2.1. Sea $\widetilde{A}_{C_i} = \bigcup_{L} \widetilde{A}_j$, el conjunto difuso del intervalo o clase C_i, $\forall j \in \{1,...,n\}$ y $L = \{x \in \Omega : L_{S_{i-1}} \leq x < L_{I_{i+1}}\}$.

Definición 3.2.2. Sea

$$\widetilde{n}_i = \#Nucleo(\widetilde{A}_{C_i}) + \int_{L_{I_i}}^{L_{S_i}} \mu_{\widetilde{A}_{C_i}}(x)\ dx,$$

la frecuencia difusa absoluta y sea

$$\widetilde{n} = \#Nucleo(\mathcal{M}) + \int_{\text{mín}(x)}^{\text{máx}(x)} \mu_{\widetilde{A}_{C_i}}(x)\ dx,$$

la frecuencia difusa total. Si $\widetilde{A}_i$ es un número convencional $\forall i \in \{1,...,n\}$, entonces

$$\int_{L_{I_i}}^{L_{S_i}} \mu_{\widetilde{A}_{C_i}}(x)\ dx = 0\ y \int_{\text{mín}(x)}^{\text{máx}(x)} \mu_{\widetilde{A}_{C_i}}(x)\ dx = 0,$$

así $\widetilde{n}_i = n_i$ y $\widetilde{n} = n$, donde n_i es la frecuencia absoluta y n el tamaño muestral.

Note que:

$$\sum_{i=1}^{k} \widetilde{n}_i = \sum_{i=1}^{k} \{\#Nucleo(\widetilde{A}_{C_i}) + \int_{L_{I_i}}^{L_{S_i}} \mu_{\widetilde{A}_{C_i}}(x)\ dx\} = \sum_{i=1}^{k} \#Nucleo(\widetilde{A}_{C_i}) + \sum_{i=1}^{k} \int_{L_{I_i}}^{L_{S_i}} \mu_{\widetilde{A}_{C_i}}(x)\ dx,$$

finalmente

$$\sum_{i=1}^{k} \widetilde{n}_i = \#Nucleo(\mathcal{M}) + \int_{\text{mín}(x)}^{\text{máx}(x)} \mu_{\widetilde{A}_{C_i}}(x)\ dx = \widetilde{n}.$$

Definición 3.2.3. Sea $\widetilde{f}_i = \frac{\widetilde{n}_i}{\widetilde{n}}$, la frecuencia relativa difusa, donde $\sum_{i=1}^{k} \widetilde{f}_i = 1$.

Conocidos los datos difusos $X_1, ..., X_n$, se tiene: $P(X \subseteq \widetilde{H}) \approx \frac{\widetilde{n}_h}{\widetilde{n}} = \widetilde{f}_h$, donde $\widetilde{n}_h$ es la frecuencia absoluta difusa limitada por la clase $\widetilde{H} = \{(x,1) : \frac{-h}{2} \leq x < \frac{h}{2}\}$. Así se obtiene la siguiente estimación de la densidad difusa:

$$\hat{\widetilde{f}}_h(x) = \frac{1}{\widetilde{n}h}\{\#Nucleo(\widetilde{A}_{C_h}) + \int_{-\frac{h}{2}}^{\frac{h}{2}} \mu_{\widetilde{A}_{C_h}}(x)\, dx\}.$$

Supongamos una muestra de observaciones difusa triangulares, con diferentes expansiones y no necesariamente simétricas, vea la Figura 1.

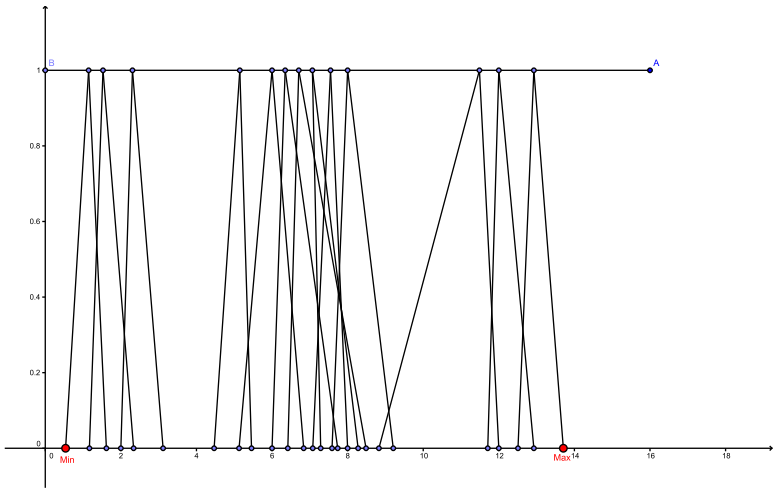

Figura 1: Esta figura presenta una muestra de números difusos triangulares no simétricos. Se destacan los valores máximo y mínimo.

La comparación es realizada utilizando el método convencional, presentado por [5], usando $k = 5$. En el Cuadro 1 son especificadas las frecuencias difusas.

Las funciones de pertenencia generadas en algunos artículos difusos con las caracteristicas de una clase difusa son denominadas funciones de pertenencia no convexas [15].

4 Kaplan-Meier difuso y comportamiento asintótico

La generalización de Kaplan-Meier cuando los tiempos de vida son considerados números difusos es presentado considerando los siguientes preliminares:

1. $\widetilde{t}_1 < \widetilde{t}_2 < ... < \widetilde{t}_k$, son k diferentes tiempos de vida difusos ordenados de acuerdo con las fallas.
2. $\widetilde{d}_j$ el número difuso de fallas en $\widetilde{t}_j, j = 1, ..., k$, y

3. $\widetilde{n}_j$ el número difuso de individuos en riesgo $\widetilde{t}_j$, esto es, individuos que no fallaron y no fueron censurados hasta el momento inmediatamente antes de $\widetilde{t}_j$.

Luego el estimador de Kaplan-Meier para tiempos de vida difusos es definido como:

$$\hat{S}(\widetilde{t}) = \prod_{j:\widetilde{t}_j<t} (\frac{\widetilde{n}_j - \widetilde{d}_j}{\widetilde{n}_j}) = \prod_{j:\widetilde{t}_j<t} (1 - \frac{\widetilde{d}_j}{\widetilde{n}_j}).$$

Teorema 4.1. *La varianza asintótica del estimador de Kaplan-Meier cuando los tiempos de vida son considerados difusos es dada por:*

$$\hat{Var}(\hat{S}(\widetilde{t})) \approx [\hat{S}(\widetilde{t})]^2 \sum_{j \,:\, \widetilde{t}_j<t} \frac{\widetilde{d}_j}{\widetilde{n}_j(\widetilde{n}_j - \widetilde{d}_j)}.$$

Demostración. Para la prueba es necesario el método delta. El método delta necesita que la función $f(x)$ sea expresada en serie de Taylor. Sea $f(x)$ una función de densidad de probabilidad de la variable aletoria X. La serie de Taylor de primer orden sobre la media es:

$$f(x) \approx f(\mu) + (x - \mu)f'(\mu). \tag{1}$$

En la ecuación (1), la varianza de la función $f(x)$ es aproximadamente igual:

$$Var(f(x)) \approx [f'(\mu)]^2 Var(x - \mu) \approx [f'(\mu)]^2\sigma^2,$$

donde σ^2 es la varianza de la variable aleatoria X. Por ejemplo, en el caso de la función logaritmo natural tenemos:

$$\ln(X) \approx \ln(\mu) + (X - \mu)\frac{1}{\mu}, \tag{2}$$

luego

$$\hat{Var}(ln(X)) \approx \frac{1}{\hat{\mu}^2}\hat{\sigma}^2. \tag{3}$$

Cuadro 1: Frecuencia Absoluta difusas.

$---\widetilde{n}_i, \quad \forall i \in \{1, ..., 5\} --$
$\int_{L_{I_1}}^{L_{S_1}} \mu_{\widetilde{A}_{C_1}}(x)\, dx = 1{,}4, \text{——} \widetilde{n}_1 = 4{,}4$
$\int_{L_{I_2}}^{L_{S_2}} \mu_{\widetilde{A}_{C_2}}(x)\, dx = 0{,}68, \text{——} \widetilde{n}_2 = 1{,}68$
$\int_{L_{I_3}}^{L_{S_3}} \mu_{\widetilde{A}_{C_3}}(x)\, dx = 2{,}32, \text{——} \widetilde{n}_3 = 8{,}32$
$\int_{L_{I_4}}^{L_{S_4}} \mu_{\widetilde{A}_{C_4}}(x)\, dx = 1{,}3, \text{——} \widetilde{n}_4 = 1{,}3$
$\int_{L_{I_5}}^{L_{S_5}} \mu_{\widetilde{A}_{C_5}}(x)\, dx = 1{,}63, \text{——} \widetilde{n}_5 = 4{,}63$

ahora, en nuestro contexto, para calcular la varianza del estimador de Kaplan-Meier basado en el método delta, tenemos:

$$\hat{S}(\widetilde{t}) = \prod_{j:\widetilde{t}_j<t} (\frac{\widetilde{n}_j - \widetilde{d}_j}{\widetilde{n}_j}),$$

luego

$$\ln(\hat{S}(\widetilde{t})) = \sum_{j:\widetilde{t}_j<t} \ln\left(1 - \frac{\widetilde{d}_j}{\widetilde{n}_j}\right). \tag{4}$$

Con la substitución de $\hat{p}_j = 1 - \frac{\widetilde{d}_j}{\widetilde{n}_j}$ en la ecuación (4) tenemos:

$$\ln(\hat{S}(\widetilde{t})) = \sum_{j:\widetilde{t}_j<t} \ln(\hat{p}_j), \tag{5}$$

usando el operador varianza en ambos lados de la ecuación (5) y asumiendo independencia entre las variables $\hat{p}_j$, tenemos:

$$\hat{Var}(\ln(\hat{S}(\widetilde{t}))) = \sum_{j:\widetilde{t}_j<t} Var(\ln(\hat{p}_j)). \tag{6}$$

Si suponemos que $\hat{p}_j$ tiene distribución Bernoulli con probabilidad p_j, luego, el estimador de p_j es $\hat{p}_j$ y como estimador de la varianza $\frac{\hat{p}_j(1-\hat{p}_j)}{\widetilde{n}_j}$, ahora usando las ecuaciones (2) y (3) tenemos:

$$\hat{Var}(\ln(\hat{p}_j)) \approx \frac{\hat{p}_j(1-\hat{p}_j)}{\hat{p}_j^2 \widetilde{n}_j},$$

luego usando el hecho que $\hat{p}_j = \frac{\widetilde{n}_j - \widetilde{d}_j}{\widetilde{n}_j}$, tenemos:

$$\hat{Var}(\ln(\hat{p}_j)) \approx \frac{\widetilde{d}_j}{\widetilde{n}_j(\widetilde{n}_j - \widetilde{d}_j)}.$$

Finalmente tenemos:

$$\hat{Var}(\ln(\hat{S}(\widetilde{t}))) \approx \sum_{j:\widetilde{t}_j<t} \frac{\widetilde{d}_j}{\widetilde{n}_j(\widetilde{n}_j - \widetilde{d}_j)}. \tag{7}$$

La expresión (7) permite obtener una estimación de la varianza del logaritmo natural de la función de sobrevivencia. Nuevamente aplicando el método delta en la función exponencial para eliminar el logaritmo natural de la expresión (7) tenemos:

$$\hat{Var}(\hat{S}(\widetilde{t})) = [\hat{S}(\widetilde{t})]^2 \sum_{j\,:\,\widetilde{t}_j<t} \frac{\widetilde{d}_j}{\widetilde{n}_j(\widetilde{n}_j - \widetilde{d}_j)}.$$

□

Basado en la definición de frecuencia difusa, también es posible proponer una expresión general para la varianza e intervalos de confianza. Asumiendo que $\hat{S}(\widetilde{t})$, para $\widetilde{t}$ fijo, también tiene distribución normal asintótica (ver Teorema 4.1.4), tiene un intervalo aproximado de $100(1-\alpha)\,\%$ de confianza para $S(\widetilde{t})$ dado por:

$$\hat{S}(\widetilde{t}) \pm z_{\alpha/2}\sqrt{\hat{Var}(\hat{S}(\widetilde{t}))}.$$

4.1 Comportamiento asintótico

Sean $\hat{S}(t)$ y $\hat{S}(\widetilde{t})$ el estimador de Kaplan-Meier convencional y para tiempos de vida difuso respectivamente, donde

$$\hat{S}(t) = \prod_{j:t_j<t} (1 - \frac{d_j}{n_j}) \text{ y } \hat{S}(\widetilde{t}) = \prod_{j:\widetilde{t}_j<t} (1 - \frac{\widetilde{d}_j}{\widetilde{n}_j}).$$

Se dice que $\widetilde{n}$ tiende a ∞, cuando $n_m \in Nucleo(\widetilde{n})$ tiende a ∞.

Teorema 4.1.1. $\|\hat{S}_n(t) - \hat{S}_n(\widetilde{t})\| \xrightarrow{P} 0$.

Demostración. Tenemos

$$\lim_{\widetilde{n}\to\infty} \|\hat{S}_n(t)\| = \lim_{\widetilde{n}\to\infty} \| \prod_{j:t_j<t} (1 - \frac{d_j}{n_j})\| = 0,$$

porque los factores $1 - \frac{d_j}{n_j} \to 0$ cuando $n \to \infty$. Por otro lado

$$\lim_{\widetilde{n}\to\infty} \|\hat{S}_n(\widetilde{t})\| = \lim_{\widetilde{n}\to\infty} \| \prod_{j:\widetilde{t}_j<t} (1 - \frac{\widetilde{d}_j}{\widetilde{n}_j})\| = 0,$$

porque los factores $1 - \frac{\widetilde{d}_j}{\widetilde{n}_j} \to 0$ cuando $\widetilde{n} \to \infty$. Además basado en el Teorema de Slutsky

$$\lim_{\widetilde{n}\to\infty} \|\hat{S}_n(t) - \hat{S}_n(\widetilde{t})\| = 0.$$

Entonces $\|\hat{S}_n(t) - \hat{S}_n(\widetilde{t})\| \xrightarrow{L_2} 0$ e $\|\hat{S}_n(t) - \hat{S}_n(\widetilde{t})\| \xrightarrow{P} 0$. □

Basado en la Definición 3.2.2 se puede escribir, $\widetilde{d}_j = d_j + \epsilon_j$ y $\widetilde{n}_j = n_j + \delta_j$, donde $\epsilon \geq 0$ y $\delta \geq 0$.

Teorema 4.1.2. *Si $\frac{d_j}{n_j} \geq \frac{\epsilon_j}{\delta_j}$, para todo j, tal que $t_j < t$. Entonces $\hat{S}(t) \leq \hat{S}(\widetilde{t})$.*

Demostración. Si $\frac{d_j}{n_j} \geq \frac{\epsilon_j}{\delta_j}$ entonces

$$d_j\delta_j \geq \epsilon_j n_j \iff d_j\delta_j - \epsilon_j n_j \geq 0 \iff d_j n_j + d_j\delta_j - \epsilon_j n_j \geq d_j n_j$$

$$\iff \frac{d_j(n_j+\delta_j)}{n_j} - \epsilon_j \geq d_j \iff \frac{d_j(n_j+\delta_j) - \epsilon_j n_j}{n_j} \geq d_j$$

$$\iff \frac{d_j(n_j+\delta_j) - \epsilon_j n_j}{n_j(n_j+\delta_j)} \geq \frac{d_j}{n_j+\delta_j} \iff \frac{d_j}{n_j} - \frac{\epsilon_j}{n_j+\delta_j} \geq \frac{d_j}{n_j+\delta_j}$$

$$\iff \frac{d_j}{n_j} \geq \frac{d_j}{n_j+\delta_j} + \frac{\epsilon_j}{n_j+\delta_j} \iff \frac{d_j}{n_j} \geq \frac{d_j+\epsilon_j}{n_j+\delta_j}$$

$$\iff 1 - \frac{d_j}{n_j} \leq 1 - \frac{d_j+\epsilon_j}{n_j+\delta_j}.$$

Finalmente $\hat{S}(t) \leq \hat{S}(\widetilde{t})$.

□

Note que si $\frac{d_j}{n_j} \leq \frac{\epsilon_j}{\delta_j}$, para todo j, tal que $t_j < t$. entonces $\hat{S}(t) \geq \hat{S}(\widetilde{t})$.

Teorema 4.1.3. $\lim\limits_{(\epsilon_j,\delta_j)\to(0,0)} \hat{S}_n(\widetilde{t}) \xrightarrow{D} N\left(E(\hat{S}(t)), Var(\hat{S}(t))\right)$.

Demostración.

$$\begin{aligned}
\lim_{(\epsilon_j,\delta_j)\to(0,0)} \hat{S}_n(\widetilde{t}) &= \lim_{(\epsilon_j,\delta_j)\to(0,0)} \prod_{j:\widetilde{t}_j<\widetilde{t}} (1-\frac{\widetilde{d}_j}{\widetilde{n}_j}) \\
&= \lim_{(\epsilon_j,\delta_j)\to(0,0)} \prod_{j:\widetilde{t}_j<\widetilde{t}} (1-\frac{d_j+\epsilon_j}{n_j+\delta_j}) \\
&= \prod_{j:t_j<t} (1-\frac{d_j}{n_j}) \\
&= \hat{S}_n(t).
\end{aligned}$$

Basado en [3] tenemos que $\hat{S}(t)$ tiene normalidad asintótica puntual en los intervalos compactos, esto es, $\sqrt{n}(\hat{S}_n(t) - S(t))$ converge en distribución para una normal, es decir, $\sqrt{n}(\hat{S}_n(t) - S(t)) \xrightarrow{D} N(0, Var(\hat{S}(t)))$. Así $\lim\limits_{(\epsilon_j,\delta_j)\to(0,0)} \hat{S}_n(t) \xrightarrow{D} N\left(E(\hat{S}(t)), Var(\hat{S}(t))\right)$. □

Es importante indicar que el estimador de Kaplan-Meier es el estimador de máxima verosimilitud, portanto todas las propiedades que eso implica deben ser consideradas [17].

Teorema 4.1.4. *Si $\hat{S}(\widetilde{t})$ es el estimador de Kaplan-Meier para tiempos de vida difusos, entonces* $\hat{S}_n(\widetilde{t}) \xrightarrow{D} N\left(E(\hat{S}(\widetilde{t})), Var(\hat{S}(\widetilde{t}))\right)$.

Demostración. Basado en la notación del Teorema 4.1 y las Expansiones de Edgeworth, tenemos:

$$\hat{S}_n(\widetilde{t}) = \Phi(p_j) + \frac{C_1(\hat{S})q_1(p_j)\phi(p_j)}{\sqrt{n}} + \frac{C_2(\hat{S})q_2(p_j) + C_3(\hat{S})q_3(p_j)\phi(p_j)}{n} + O(n^{-\frac{3}{2}}),$$

donde $E(\hat{S}(\widetilde{t})) = \mu$; $C_1(\hat{S}) = \frac{E\left(\hat{S}(\widetilde{t})-\mu\right)^3}{6\sigma^3}$; $C_2(\hat{S}) = \frac{\frac{E\left(\hat{S}(\widetilde{t})-\mu\right)^4}{\sigma^4}-3}{24}$; $C_3(\hat{S}) = \frac{C_1^2(\hat{S})}{72}$; $q_1(p_j) = (1-p_j^2)$; $q_2(p_j) = 3p_j - p_j^3$ y $q_3(p_j) = 10p_j^3 - 15p_j - p_j^5$. Observe que cuando $\widetilde{n} \to \infty$ se tiene que $p_j \to 0$ (ver Teorema 4.1), entonces $q_1(p_j) = (1-p_j^2) \to 1$; $q_2(p_j) = 3p_j - p_j^3 \to 0$ y $q_3(p_j) = 10p_j^3 - 15p_j - p_j^5 \to 0$, luego se puede escribir:

$$\hat{S}_n(\widetilde{t}) = \Phi(p_j) + \frac{C_1(\hat{S})\phi(p_j)}{\sqrt{n}} + O(n^{-\frac{3}{2}}).$$

El análisis del comportamiento asintótico de $C_1(\hat{S}) = \frac{E\left(\hat{S}(\widetilde{t})-\mu\right)^3}{6\sigma^3}$ se basa en el método Delta (ver Teorema 4.1), se tiene que $C_1(\hat{S}) = \frac{E\left(\hat{S}(\widetilde{t})-\mu\right)^3}{6\sigma^3} \to 0$ cuando $\widetilde{n} \to \infty$. Finalmente $\hat{S}_n(\widetilde{t}) \xrightarrow{D} N\left(E(\hat{S}(\widetilde{t})), Var(\hat{S}(\widetilde{t}))\right)$.

□

5 Aplicación

Los tiempos de vida para el siguiente problema, no son difusos, sin embargo para efectos de la propuesta, serán asumidos con esta estructura (ver Tabla 3). Se realizó un estudio clínico aleatorio para investigar el efecto de la terapia con esteroides en el tratamiento de la hepatitis viral aguda [12]. Veintinueve pacientes con la enfermedad fueron aleatorizados para recibir un elemento inerte o tratamiento con esteroides. Cada paciente fue seguido por 16 semanas o hasta la muerte. Este conjunto de datos es estudiado por [5].

El Cuadro 2 muestra la función de sobrevivencia estimada basado en las estimaciones de Kaplan-Meier convencional [5].

Cuadro 2: Kaplan-Meier para el grupo de esteroides.

t_j	Intervalo	d_j	n_j	$\hat{S}(t_{j+})$
0	[0,1)	0	14	1,000
1	[1,5)	3	14	0,786
5	[5,7)	1	9	0,698
7	[7,8)	1	8	0,611
8	[8,10)	1	7	0,524
10	[10,16)	1	6	0,437

Basado en el mismo problema descrito en el Cuadro 2, se asume una representación difusa para cada tiempo de falla, como es mostrado en la Figura 1. Observe que las funciones de pertenencia son triangulares y no se asume simetría. Basado en las definiciones 2.6, 3.2 y 4, el estimador de Kaplan-Meier para tiempos de vida difusos es presentado en el Cuadro 3.

Cuadro 3: Estimaciones de Kaplan-Meier para el grupo de esteroides con tiempos difusos.

$\widetilde{t}_j$	Intervalo	$\widetilde{d}_j$	$\widetilde{n}_j$	$\widetilde{\hat{S}}(t_{j+})$
(0,0,0)	[0,1)	0.35	16.68	1,000
(0.3,1,1,3)(0.5,1,1.3)(0.7,1,1.7)	[1,5)	3.7	16.33	0,773
(4.3,5,5.3)	[5,7)	1.4	12.63	0,687
(6.5,7,8)	[7,8)	1.78	11.23	0,578
(6.7,8,8.3)	[8,10)	1.3	9.45	0,498
(9.7,10,10.3)	[10,16)	1.15	8.15	0,427

El Cuadro 4 muestra una pequeña comparación entre la determinación de la varianza y los intervalos de confianza en dos situaciones: de manera convencional y considerando los tiempos de vida como números difusos. Se recuerda que la notación (a_l, a_m, a_r) representa un número difuso tipo LR. Se observa que el Cuadro 4, la varianza disminuye y las amplitudes de los intervalos de confianza disminuyen, mostrando el efecto de la información adicional proporcionada por la consideración de los tiempos de falla como números difusos.

Cuadro 4: Estimaciones de Kaplan-Meier para los grupos de esteroides de manera convencional y con tiempos de vida difusos.

t_j	Varianza	Intervalo de confianza
$t_j = 5$	0.0163	$0{,}698 \pm 1{,}96\sqrt{0{,}0163} = (0{,}45; 0{,}95)$
$\widetilde{t}_j = (4{,}3, 5, 5{,}3)$	0.0137	$0{,}687 \pm 1{,}96\sqrt{0{,}0137} = (0{,}458; 0{,}916)$

6 Conclusiones

En este trabajo es generalizado el estimador de Kaplan-Meier, cuando los tiempos de falla son considerados difusos. Es presentado un nuevo concepto de aproximaciones para las frecuencias relativas difusas y sus propiedades asintóticas. Se promueve el uso de estructuras numéricas en coherencia a la naturaleza humana, superando problemas de idealizaciones imprecisas. Uno de sus efectos es una disminución de la variabilidad y por tanto, una disminución en la amplitud de los intervalos de confianza. Como trabajos futuros se pretende estudiar el comportamiento asintótico del estimador de Kaplan-Meier.

7 Agradecimientos

J.A. González agradece el financiamiento entregado por la Dirección General de Investigación (fondo 2017) y Unidad de Soporte Estadístico para la Investigación.

Referencias

[1] Amato, D. A., *A generalized kaplan-meier estimator for heterogenous populations*, Communications in Statistics-Theory and Methods, **17**(1)(1988), 263-286.

[2] Bitouze, D., Laurent, B. and Massart, P., *A dvoretzky – kiefer – wolfowitz type inequality for the kaplan–meier estimator*, In Annales de l'Institut Henri Poincare (B) Probability and Statistics, **35**(1)(1999), 735-763.

[3] Breslow, N., *Covariance analysis of censored survival data*, Biometrics, **3**(1)(1974), 89-99.

[4] Cai, Z., *Asymptotic properties of kaplan-meier estimator for censored dependent data*, Statistics and probability letters, **37**(4)(1998), 381-389.

[5] Colosimo, E. A. and Giolo, S. R., *Análise de sobrevivencia aplicada*. In ABE-Projeto Fisher, Edgard Blucher, Sao Paulo, Brasil, 2006.

[6] Dabrowska, D. M., *Kaplan-meier estimate on the plane*, The Annals of Statistics, **7**(4) (1988), 1475-1489.

[7] Dabrowska, D. M., *Kaplan-meier estimate on the plane: weak convergence, lil, and the bootstrap*, Journal of Multivariate analysis, **29**(2)(1989), 308-325.

[8] Dubois, D. and Prade, H, (1978). *Operations on fuzzy numbers*, International Journal of systems science, **9**(6)(1978), 613-626.

[9] Egusa, Y., Akahori, H., Morimura, A. and Wakami, N., *An application of fuzzy set theory for an electronic video camera image stabilizer*, Fuzzy Systems, IEEE Transactions on, **3**(3)(1995), 351-356.

[10] Goguen, J. A., *L-fuzzy sets*, Journal of mathematical analysis and applications, **18**(1)(1967), 145-174.

[11] González C., J. A., *Estatística e a teoria de conjuntos fuzzy*, Tese Doutorado Universidade Estadual de Campinas, Brasil , **1**(1)(2015), 1-77.

[12] Gregory, P. B., Knauer, C. M., Kempson, R. L. and Miller, R., *Steroid therapy in severe viral hepatitis: A double-blind, randomized trial of methyl-prednisolone versus placebo*, New England Journal of Medicine, **294**(13)(1976), 681-687.

[13] Huang, H.-Z., Zuo, M. J. and Sun, Z.-Q., *Bayesian reliability analysis for fuzzy lifetime data*, Fuzzy Sets and Systems, **157**(12)(2006), 1674-1686.

[14] Meier-Kriesche, H.-U., Schold, J. D. and Kaplan, B., *Long-term renal allograft survival: Have we made significant progress or is it time to rethink our analytic and therapeutic strategies?*, American Journal of Transplantation, **4**(8)(2004), 1289-1295.

[15] Nasseri, S., Taleshian, F., Alizadeh, Z. and Vahidi, J., *A new method for ordering lr fuzzy number*, The Journal of Mathematics and Computer Science, **4**(3)(2012), 283-294.

[16] Stute, W., *The jackknife estimate of variance of a kaplan-meier integral*, The Annals of Statistics, **24**(6)(1996), 2679-2704.

[17] Van der Vaart, A. W., *Asymptotic statistics*. Cambridge university press, 2000.

[18] Viertl, R., *On reliability estimation based on fuzzy lifetime data*, Journal of Statistical Planning and Inference, **139**(5)(2009), 1750-1755.

[19] Wang, L.-X., *A course in fuzzy systems*. Prentice-Hall press, USA, 1999.

[20] Xie, J. and Liu, C., *Adjusted kaplan–meier estimator and log-rank test with inverse probability of treatment weighting for survival data*, Statistics in medicine, **24**(20)(2005), 3089-3110.

[21] Zadeh, L. A., *Fuzzy sets*, Information and control , **8**(3)(1965), 338-353.

[22] Zadeh, L. A., *The concept of a linguistic variable and its application to approximate reasoningi*, Information sciences, **8**(3)(1975), 199-249.