P. CH. MAHALANOBIS Y LAS APLICACIONES DE SU DISTANCIA ESTADISTICA

Mtra. María Teresa Escobedo Portillo[1] y PhD Jorge A. Salas Plata Mendoza[2]

RESUMEN

El objetivo de este texto es introducir al lector en el tema de la Distancia de Mahalanobis y subrayar las principales aplicaciones de este concepto estadístico que ayuda, de manera directa o indirecta, al acopio, organización y análisis de datos numéricos y a la solución de problemas de diseño de experimentos y toma de decisiones. Como objetivo secundario, se hace un bosquejo de la vida y obra de este científico. Se lleva a cabo una revisión de la literatura para mostrar una serie de ejemplos de aplicación que ilustran las contribuciones de P. Ch. Mahalanobis a diferentes áreas del conocimiento y para destacar los momentos importantes de la vida de este investigador. Se logra conjuntar en un solo documento tanto los principales descubrimientos como el esbozo de la biografía del investigador. Se concluye que el enfoque estadístico de este intelectual ha posibilitado, en la actualidad, el uso de su distancia a problemas en los que se busca conocer no sólo la distancia entre las variables, sino su correlación, superando las limitaciones de la Distancia Euclidanea. Estas contribuciones, apoyadas en la búsqueda de aplicaciones prácticas, reflejan en gran medida, las preocupaciones intelectuales, experiencias profesionales e influencias familiares de Mahalanobis.

Palabras clave: Distancia de Mahalanobis, datos numéricos, diseño experimental.

1. INTRODUCCIÓN

En esta sección se explica la importancia el concepto de la Distancia de Mahalanobis y se ofrecen ejemplos de aplicación. Se destacan aspectos relevantes de la vida del investigador, asociadas a su producción intelectual.[12]

1.1 Planteamiento del problema

La mayoría de los escritos relativos a la Distancia de Mahalanobis ofrecen pocos ejemplo de aplicación, por lo que el lector acaso tiene una visión general del impacto de este concepto estadístico a diferentes áreas del conocimiento; de igual manera, los artículos hacen poca referencia a la vida del autor asociada a sus descubrimientos, por lo que se carece de una visión integral de dichos hallazgos.

1.2 Revisión de la literatura

Prasanta Chandra Mahalanobis, quien nació el 29 de junio de 1893 y murió el 28 de junio de 1972, fue un científico de la India que destacó en estadística aplicada. Su padre, Prabodh Chandra, fue un miembro activo del movimiento religioso Bramho Samaj. Su madre, Nirodbasin, perteneció a una familia de gran tradición académica. Mahalanobis se graduó en física en 1912 por la Universidad presidencial de Kolkata, y terminó sus estudios en el King's College de Cambdridge, para posteriormente volver a Calcuta. Fue conocido popularmente como físico por formación, estadístico por instinto y planeador por convicción, ya que usó ideas sencillas para desarrollar modelos econométricos en países

[1] Instituto de Ingeniería y Tecnología. UACJ. mtescobe@uacj.mx
[2] Instituto de Ingeniería y Tecnología. UACJ. jsalas@uacj.mx

en vías de desarrollo. El avizoró que la estadística, una ciencia nueva relacionada con las mediciones, tenía un amplio potencial de aplicaciones. Mahalabonis desarrolló el estadístico D^2, conocido como o la "Distancia de Mahalanobis" (Rao, 2005). Realizó trabajos pioneros en el estudio de las variaciones antropométricas en la India, fundó el Instituto Indio de Estadística, y contribuyó al campo de las encuestas a gran escala. Inspirado por la revista científica Biometrika y por Acharya Brajendranath Seal, empezó a trabajar en estadística analizando resultados de exámenes universitarios, medidas antropométricas de anglo-indios de Calcuta y problemas metereológicos. También trabajó como metereólogo durante algún tiempo. En 1924, mientras trabajaba en la probabilidad de error de los resultados de los experimentos en agricultura, conoció a Ronald Fisher, con quien estableció una amistad que se mantendría durante toda su vida. También trabajó en modelos para prevenir inundaciones. Mahalanobis llevó a cabo tres contribuciones notables en técnicas de muestreo: proyectos piloto, diseño de proyectos óptimos e interpretación de redes de muestras. Un proyecto piloto suministra información básica con relación a costos operativos y la incertidumbre de las variables de dicho proyecto. La precisión del muestreo depende, de acuerdo con este investigador, de tres aspectos: a) el tamaño óptimo de las unidades de muestreo, b) el total de las unidades de muestreo que deben usarse para obtener un cierto grado de precisión en los estimados finales y c) la mejor manera de distribuir las unidades de muestreo en los distritos, regiones o zonas cubiertas por el estudio (Rao, 2005).

Mahalanobis también promovió en su país técnicas de Control Estadístico de la Calidad e Investigación de Operaciones en el sector industrial, para incrementar la productividad y mejorar la calidad de los bienes manufacturados. Mahalanobis mostró interés por los logros culturales y fue secretario de Rabindranath Tagore, particularmente durante sus viajes al extranjero. Recibió el premio Padma Vidhushan, uno de los premios más reputados de la India, por sus contribuciones a la ciencia y sus servicios al país. En sus últimos años continuó su labor investigadora y desempeño los cargos de Secretario y Director del Instituto Indio de Estadística y Consejero Honorífico de Estadística del Gabinete de Gobierno de la India. Obtuvo los siguientes premios:

- Medalla Weldon de la Universidad de Oxford (1944)
- Socio de la Royal Society, de Londres (1945)
- Presidente Honorífico del Instituto Internacional de Estadística (1957)
- Padma Vibhushan (1968) (Wikipedia, 2008)

1.2.1 La Distancia de Mahalanobis

En estadística, la Distancia de Mahalanobis es una medida de distancia introducida por Mahalanobis en 1936. Su utilidad radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales. Se diferencia de la distancia euclídea en que tiene en cuenta la correlación entre las variables aleatorias.

Para entender el concepto de la Distancia de Mahalanobis, es necesario recurrir a varias definiciones y descripciones de tipos de variables, ya que se trata de encontrar las correlaciones entre dichas variables.

1.2.2 Tipos de Variables

- Variable aleatoria: variable que cuantifica los resultados de un experimento aleatorio. Variable que toma diferentes valores como resultado de un experimento aleatorio. Categoría

cuantificable que puede tomar diferentes valores cada vez que sucede un experimento o suceso, el valor sólo se conocerá deterministamente una vez acaecido el suceso. En estadística y teoría de la probabilidad, una variable aleatoria se define como el resultado numérico de un experimento aleatorio. Matemáticamente, es una aplicación $X : \Omega \to \mathbb{R}$ que da un valor numérico, en el conjunto de los reales, a cada suceso en el espacio Ω de los resultados posibles del experimento. Dada una variable aleatoria *X,* se pueden calcular estimadores estadísticos diferentes como la media (media aritmética, media geométrica, media ponderada) y el valor esperado y la varianza de la distribución de probabilidad de *X*. Para ilustrar el concepto de variable aleatoria se presenta el siguiente ejemplo:

Para ilustrar esta noción, suponga que lanzamos dos monedas al aire; sea X la variable aleatoria que identifica el número de caras obtenidas en el lanzamiento.

X: Número de caras obtenidas en el lanzamiento.
Ω = { cc, cs, sc, ss } (c identifica una cara, s un sello)
R_X = { 0, 1, 2 } (Recorrido de X)

Entonces, asociando a Omega con el Recorrido de X, tenemos que:

$$X : \Omega \to R_X$$
$$cc \to 2$$
$$cs, sc \to 1$$
$$ss \to 0$$

- Variable aleatoria discreta: variable que toma un número finito o infinito de valores numerables.
- Variable aleatoria continua: variable que toma un valor infinito de valores no numerables. Variable aleatoria que puede tomar cualquier valor dentro de un intervalo dado de valores.

1.2.3 Otras definiciones

- Distribución de probabilidades: modelo teórico que describe la forma en que varían los resultados de un experimento aleatorio. Lista de los resultados de un experimento con las probabilidades que se esperarían ver asociadas con cada resultado.
- Función de probabilidad: función que asigna probabilidades a cada uno de los valores de una variable aleatoria discreta.
- Función de densidad de probabilidad: función que mide la concentración de probabilidad alrededor de los valores de una variable aleatoria continua.
- Función de distribución: función que acumula probabilidades asociadas a una variable aleatoria.
- Valor esperado o esperanza matemática: operador matemático que caracteriza la posición de la distribución de probabilidades. Media ponderada de los resultados de un experimento (Wikipedia, 2008).

1.2.4 Distancia

Se denomina distancia a la longitud del camino más corto entre dos entidades. Desde un punto de vista formal, para un conjunto de elementos *X,* se define distancia como cualquier función binaria *d*(*a*,*b*) de $X \times X$ en $\mathbb{R}$ que verifique las siguientes condiciones:

- No negatividad:

$$d(a,b) \geq 0 \ \forall a,b \in X$$

- Simetricidad:

$$d(a,b) = d(b,a) \ \forall a,b \in X$$

- Desigualdad triangular:

$$d(a,b) \leq d(a,c) + d(c,b) \ \forall a,b,c \in X$$

1.2.5 Distancia (geometría)

Se denomina distancia euclídea entre dos puntos $A(x_1,y_1)$ y $B(x_2,y_2)$ a la longitud del segmento de recta que tiene por extremos A y B. Se expresa matemáticamente como:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

La distancia entre un punto P y una recta R es la longitud del camino más corto que une el punto $P(x_1,y_1)$ con la recta R: $Ax + By + C = 0$. Matemáticamente se expresa como:

$$d = \frac{Ax_1 + By_1 + C}{\sqrt{A^2 + B^2}}$$

La distancia entre dos rectas paralelas es la longitud del camino más corto entre una de ellas y un punto cualquiera de la otra. La distancia entre un punto P y un plano L es la longitud del camino más corto entre el punto $P(x_1,y_1,z_1)$ y el plano $L = Ax + By + Cz + D$. Matemáticamente se expresa como:

$$d = \frac{Ax_1 + By_1 + Cz_1 + D}{\sqrt{A^2 + B^2 + C^2}}$$

Formalmente, la distancia de Mahalanobis entre dos variables aleatorias con la misma distribución de probabilidad $\vec{x}$ y $\vec{y}$ con matriz de covarianza Σ se define como:

$$d_m(\vec{x},\vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}.$$

1.2.6 Propiedades de la distancia de Mahalanobis

La distancia de Mahalanobis cumple las siguientes propiedades, necesarias para ser una distancia:

- Semipositividad:

$d(a,b) \geq 0 \ \forall a,b \in X$ y además
$d(a,b) = 0 \ si \ a = b$

Es decir, la distancia entre dos puntos de las mismas coordenadas es cero, y si tienen coordenadas distintas la distancia es positiva, pero nunca negativa.

- Simetricidad:

$$d(a,b) = d(b,a) \ \forall a,b \in X$$

Intuitivamente, la distancia ente *a* y *b* es la misma que entre *b* y *a*.

- Desigualdad triangular:

$$d(a,b) \leq d(a,c) + d(c,b) \ \forall a,b,c \in X$$

(Wikipedia, 2008)

1.3 Objetivo

El objetivo de este documento es introducir al lector en el tema de la Distancia de Mahalanobis y destacar las principales aplicaciones de este concepto estadístico que ayuda, de manera directa o indirecta, al acopio, organización y análisis de datos numéricos y a la solución de problemas de diseño de experimentos y toma de decisiones. Como

objetivo secundario, se hace un bosquejo de la vida y obra de este científico.

2. METODOLOGIA

Se lleva a cabo una revisión de la literatura para mostrar una serie de ejemplos de aplicación que ilustran las contribuciones de P. Ch. Mahalanobis a diferentes áreas del conocimiento y para destacar los momentos importantes de la vida de este investigador.

Para entender la utilidad de la distancia de Mahalanobis, se puede considerar el siguiente ejemplo práctico: Un pescador quiere poder medir la similitud entre dos salmones, porque quiere clasificarlos en dos tipos para su venta y poder así vender los grandes más caros. Para cada salmón mide su anchura y su longitud. Con estos datos construye un vector $\vec{x}_i = (x_{1i}, x_{2i})$ para cada salmón *i*.

La longitud de los salmones pescados es una variable aleatoria que toma valores entre 50 y 100 cm, mientras que su anchura está entre 10 y 20 cm. Si el pescador usase la distancia euclidea:

$$d_e(\vec{x_1}\vec{x_2}) = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2}$$

O en notación vectorial:

$$d_e(\vec{x_1}\vec{x_2}) = \sqrt{(\vec{x}_1 - \vec{x}_2)^T(\vec{x}_1 - \vec{x}_2)}$$

Al ser las diferencias de anchura menos grandes que las de longitud, les estará dando menos importancia. Por esta razón, el pescador decide incorporar la estadística de los datos a la medida de distancia, ponderando según su varianza; las variables con menos varianza tendrán más importancia que las de mayor varianza. De esta forma pretende igualar la importancia de la anchura y la longitud en el resultado final. La expresión quedaría:

$$d_2(\vec{x_1}\vec{x_2}) = \sqrt{\left(\frac{(x_{11} - x_{12})}{\sigma_1}\right)^2 + \left(\frac{(x_{21} - x_{22})}{\sigma_2}\right)^2}$$

Donde σ_i es la varianza de la componente *i* de los vectores de medidas.

O en notación vectorial:

$$d_e(\vec{x_1}\vec{x_2}) = \sqrt{(\vec{x}_1 - \vec{x}_2)^T S^{-1}(\vec{x}_1 - \vec{x}_2)}$$

Donde *S* es una matriz diagonal cuyos elementos en la diagonal $s_{ii} = \sigma_i$

Pero la expresión anterior tiene un problema, y es que la longitud y anchura de los salmones no son independientes; es decir, la anchura depende en cierta forma de la longitud, pues es más probable que un salmón largo sea también más ancho. Para incorporar la dependencia entre las dos variables, el pescador puede sustituir la matriz diagonal *S* por la matriz de covarianza Σ:

$$d_m(\vec{x_1}\vec{x_2}) = \sqrt{(\vec{x}_1 - \vec{x}_2)^T \Sigma^{-1}(\vec{x}_1 - \vec{x}_2)}$$

Que es la distancia de Mahalanobis (Pérez, 2007).

2.1 Otras aplicaciones de actualidad

2.1.1 Distancia generalizada de Malahanobis para una mezcla de datos

León y Carriere (2005), desarrollaron una generalización de la distancia de Mahalanobis para una mezcla de variables nominales, ordinarias y continuas. Los autores parten del supuesto de que la distancia estadística entre

poblaciones está presente en muchas de las técnicas de análisis multivariado, con variables continuas. Sin embargo, tratándose de una mezcla de variables discretas y continuas, no existe un modelo de distancia estándar. Esta distancia debe considerar, de acuerdo con los autores, no sólo los diferentes niveles de medición de las variables, sino los diversos tipos de asociaciones entre las variables.

2.1.2 Método de monitoreo de la fatiga y ruptura de herramientas de corte

Ji, et al (2002), desarrollaron un nuevo método para obtener la distancia de Mahalanobis aplicada a la fatiga y ruptura en herramientas de corte en la industria, con base en imágenes. Los autores establecen que, en comparación con la distancia Euclidiana, la de Mahalanobis tiene una relación no sólo con la distribución de cada componente del sistema o conjunto, sino con la distribución de los píxeles de la imagen dentro de cada componente de dicho conjunto. De acuerdo con los autores, este método tiene una mayor sensibilidad en la inspección de las condiciones de abrasión en las herramientas, en comparación con los métodos tradicionales.

2.1.3 Un estudio de valores de parámetro para un clasificador fuzzy de la distancia de Mahalanobis.

Deer y Eklund (2000) encontraron una aplicación en patrones de reconocimiento del parámetro fuzzy de partición *c,* en los píxeles de imágenes de sensores remotos. La adopción de la clasificación fuzzy está motivada por la presencia de problemas de mezclas de píxeles en imágenes de sensores remotos. Se han encontrado muy buenas aplicaciones en la industria forestal como los ejemplos que abordan estos autores en su artículo.

2.1.4 Modelación bioclimática con énfasis den la distancia de Mahalanobis

Farberr y Kadmon (2003), introducen en este artículo el concepto de la Distancia de Mahalanobis a la modelación bioclimática. Ellos afirman que la envolvente climática definida a través de la distancia de Mahalanobis suministra predicciones más precisas de las distribuciones de especies vivientes que las envolventes de los métodos rectilíneos tradicionales. Los autores señalan que un modelo predictivo que se ha aplicado en varios estudios científicos y para propósitos prácticos, es el llamado Modelo de Envolvente Climático (MEC). Sin embargo, el uso del MEC tiene las siguientes limitaciones: a) la incapacidad para hacer frente a las correlaciones e interacciones entre factores climáticos; b) asignación de igual adecuabilidad en combinaciones climáticas dentro de las fronteras de la envolvente climática y c) sensibilidad a los datos extremos.

2.1.5 Identificación de materiales por medio de la distancia de Mahalanobis

En este artículo, Bacci et al (2004), clasificaron muestras de prueba en el espacio principal de los componentes de un espacio maestral mediante el uso de la Distancia de Mahalanobis. Lo anterior, en un estudio de reluctancia espectroscópica en capas de pintura.

2.1.6 El sistema Mahalanobis-Taguchi (SMT)

El SMT incorpora los tres métodos estratégicos del diseño de un sistema de información. La primera estrategia introduce sólo una medida de escala en cualquier espacio multidimensional, usando la DM a cualquier subconjunto del espacio seleccionado como uniforme y calcula la distancia de la norma con relación a la distancia de otros miembros. La segunda estrategia consiste en utilizar la

relación Signal-to-Noise (SN) de la distancia, con relación al número del espacio conocido como valor real de la clasificación real. La tercera estrategia consiste en optimizar todos los factores de la información para mejorar la relación SN con un arreglo ortogonal. El SMT es una medida o herramienta de evaluación que se usa para reconocer un patrón a partir de datos multidimensionales. En el SMT, la calidad de las mediciones se evalúa con la relación SN (Taguchi, 2001).

3. DISCUSION

La distancia de Mahalanobis proporciona un método muy poderoso para saber si un determinado conjunto de condiciones similares es en realidad un conjunto de condiciones ideales, y es muy útil para identificar qué partes de un escenario son las más parecidas a las de un escenario "ideal". Por ejemplo, en Biología se puede decir que un territorio "ideal" es aquel que mejor se ajusta al nicho de ciertas especies silvestres. A través de la observación, uno puede reconocer que las especies silvestres se ubican dentro de un rango de elevaciones, pendientes específicas o de cierta densidad de vegetación. Por medio de la distancia de Mahalanobis se puede describir cuantitativamente el territorio de un animal en términos de cuán similar es a la elevación, pendiente, y densidad de vegetación ideales para dicho animal. Las distancias de Mahalanobis se basan tanto en la media y varianza de las variables predictoras, como en la matriz de covarianza de todas las variables, y por lo tanto utiliza como ventaja la covarianza entre variables (Jenness, 2008). Difiere de la distancia euclidiana en que toma en consideración la correlación del conjunto de datos en escala invariante, es decir que no depende de una escala única de medidas como lo es la edad, el peso y la altura (Data Mining, 2008). La distancia Euclidiana es insensible a las variables correlacionadas. Tomemos como ejemplo un conjunto se cinco variables, donde una sea copia fiel de otra. La copia es su gemela y por tanto completamente correlacionada. La distancia Euclidiana no tiene manera de tomar en consideración que la copia no aporta información nueva y por tanto, en los cálculos, hará pesar más esta variable que las otras. En este artículo se logró conjuntar tanto los principales descubrimientos como el esbozo de la biografía del investigador. El enfoque estadístico de este intelectual ha posibilitado, en la actualidad, el uso de su distancia a problemas en los que se busca conocer no sólo la distancia entre las variables, sino su correlación, superando las limitaciones de la Distancia Euclidanea. Estas contribuciones, apoyadas en la búsqueda de aplicaciones prácticas, reflejan en gran medida, las preocupaciones intelectuales, experiencias profesionales e influencias familiares de Mahalanobis.

4. REFERENCIAS

Bacci, M. et al. 2004. *Non-invasive fiber optic Fourier transform-infrared reflectance spectroscopy on painted layers. Identification of materials by means of principal component analysis and Mahalabonis distance*. Analytica Chimica Acta. 446 15-21.

Data Mining in MATLAB. *Mahalanobis Distance*. [En línea]. Disponible en: <http://matlabdatamining.blogspot.com/2006/11/mahalanobis-distance.html> [Consulta 20 de noviembre de 2008]

Deer P. J., and P. Eklund. 2000. *A study parameter values for a Mahalanobis Distance fuzzy classifier*. *Fuzzy Sets and Systems*. 137, 191-213.

Farber, O. y R. Kadmon. 2003. *Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance*. Ecological Modeling. 160, 115-130.

Jenness Enterprises. *Arc View Extension: Mahalanobis Description*. [En línea]. Disponible en <http://www.jennessent.com/arcview/mahalanobis_description.htm> [Consulta 20 de noviembre de 2008]

Ji, S.M., et al. 2002. *Method of monitoring wearing and breakage states of cutting tools based on Mahalabonis*

distance features. Journal of Materials Processing Technology. 129, 114-117.

Leon, A.R. y K.C., Carriere. 2005. *A generalized Mahalanobis distance for mixed data*. Journal of Multivariate Analysis. 2005. 92, 174-185.

Pérez López, César. 2007. *Métodos estadísticos avanzados con SPSS*. ITES-Paraninfo.

Rao, C. R. Mahalanobis, y Prasanta Chandra. 2005. *Enciclopedia of Social Measurements*. 2, 609-615.

Taguchi, Genichi, S. Chowdhury y Y. Wu. 2001. *The Mahalanobis-Taguchi System*. New York: American Supplier Institute, Inc. McGraw-Hill.

Wikipedia. La enciclopedia libre. *Variable aleatoria* [En línea]. La enciclopedia libre. Disponible en: <http://es.wikipedia.org/wiki/Variable_aleatoria> [Consulta 20 de noviembre de 2008]

Wikipedia. La enciclopedia libre. *Prasanta Chandra Mahalanobis* [En línea]. Disponible en: <http://es.wikipedia.org/wiki/Prasanta_Chandra_Mahalanobis> [Consulta 20 de noviembre de 2008]

Wikipedia. La enciclopedia libre. *Distancia de Mahalanobis* [En línea]. La enciclopedia libre. Disponible en: <http://es.wikipedia.org/wiki/Distancia_de_Mahalanobis> [Consulta 20 de noviembre de 2008]