El Concepto de Distancia y su Aplicación en Estadística Multivariada

Fabián M. Hernández Arellano **Millward Brown México**

0. Introducción

El propósito de este artículo es presentar y discutir el concepto de *distancia* y posteriormente examinar cómo es aplicado en Estadística Multivariada. En la primera sección del artículo se revisa el concepto *usual* de distancia y se formaliza la idea desde el punto de vista matemático al introducir la Distancia Euclidiana. En la sección 2 revisamos los supuestos involucrados en la Distancia Euclidiana y las consecuencias de que no se cumplan dichos supuestos. En la tercera sección consideramos alternativas a la Distancia Euclidiana y su aplicación en Estadística Multivariada.

1. Distancia

El significado más común en la vida cotidiana de la palabra *distancia* es el de *lejanía*. Por ejemplo, la distancia de México a Acapulco es de 400 km. La lejanía es determinada por un número que tiene *unidades*, kilómetros en el ejemplo anterior.

La palabra ha sido utilizada coloquialmente para indicar la diferencia notable entre unas cosas y otras (*la distancia entre las habilidades de escritor de Miguel de Cervantes Saavedra y las de Corín Tellado es enorme*) o bien, para indicar que dos personas han dejado de tratarse de manera cercana (*mi hermano y yo nos hemos distanciado*). Algunas personas incluso utilizan la palabra distancia para denotar el tiempo que tomó realizar una actividad (la distancia de México a Chicago en avión es de tres horas y media). Para incrementar la confusión, se ha divulgado la oración: *la distancia más corta entre dos puntos es la línea recta* (que es una barbaridad ya que lo que se quiere decir es que la *trayectoria* más corta entre dos puntos es la línea recta).

Desde el punto el vista matemático, una distancia es una función (una regla de asociación) que a un par de objetos le asocia un número real no negativo y que satisface tres condiciones. Es usual denotar la función con la letra d. Si los objetos pertenecen a un conjunto **A**, el conjunto de todos los pares ordenados de objetos es **A**x**A**, el Producto Cartesiano de **A** con **A**. De esta manera, el *dominio* de d es **A**x**A** y el *recorrido* es [0,∞), esto es,

$$d: \mathbf{A}x\mathbf{A} \rightarrow [0,\infty)$$

Así, si **x** y **y** son elementos de **A** (**x**,**y**∈ A) la distancia de **x** a **y** es denotada por d(**x**,**y**). La función d debe satisfacer las tres condiciones siguientes:

i) d(**x**,**y**) = d(**y**,**x**), esto es, d es *simétrica* (la distancia de **x** a **y** es la misma que la distancia de **y** a **x**).

ii) d(**x**,**y**) ≥ 0 y es igual a cero si y sólo si **x**=**y** (la distancia es un número real no negativo y vale cero únicamente cuando **x** y **y** son el mismo objeto).

iii) d(**x**,**y**) ≤ d(**x**,**z**) + d(**z**,**y**) para cualquier **z** elemento de A, esto es, se cumple la Desigualdad del Triángulo (la distancia de **x** a **y** es menor o igual que la suma de la distancia de **x** a **z** más la distancia de **z** a **y**. El lector puede visualizar esta propiedad al pensar que **x**, **y**, **z** definen un triángulo en el plano).

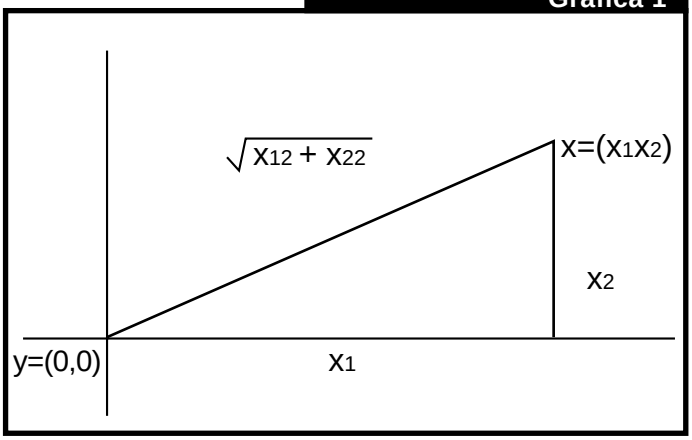
Supongamos que los objetos de interés son puntos en el espacio de n dimensiones, esto es:
x = (x_1, x_2, …, x_n) y **y** = (y_1,y_2, …, y_n)
donde x_1, x_2, …, x_n, y_1, y_2, …, y_n son números reales. La *Distancia Euclidiana* entre **x** y **y** está dada por la fórmula:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} \qquad (1)$$

Es posible verificar que d definida según la fórmula (1) satisface las tres condiciones de una distancia (el lector puede consultar el libro de Kolmogorov y Fomin (1970) página 38).

Cuando n=2, la Distancia Euclidiana tiene su razón de ser en el célebre *Teorema de Pitágoras*. De hecho si **x**=(x_1,x_2) y **y**=(0,0) al graficar los puntos en un sistema Cartesiano de dos dimensiones, se ve que la distancia de **x** a **y,** d(**x,y**), es la hipotenusa de un triángulo rectángulo de catetos x_1 y x_2.Vea la Gráfica 1.

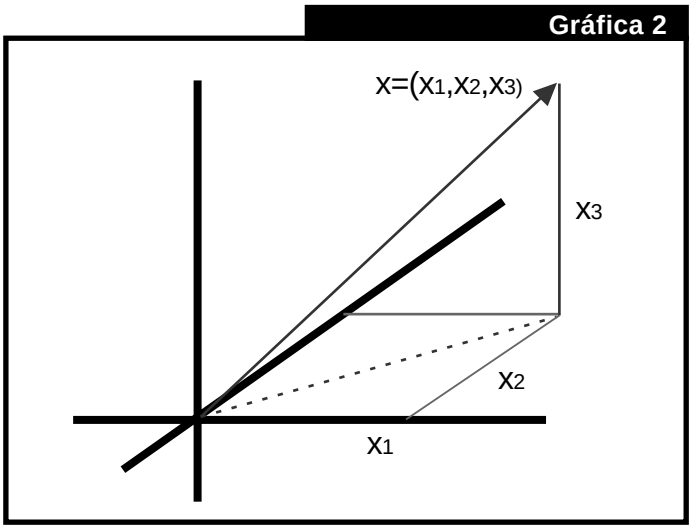
Gráfica 1



Cuando n=3, la Distancia Euclidiana tiene su fundamento en una generalización del Teorema de Pitágoras. Vea la Gráfica 2.

Los casos particulares n=2 y n=3, sugieren que la Distancia Euclidiana es la *Distancia en Línea Recta*. Para un número arbitrario de dimensiones, la Distancia Euclidiana apela a la generalización basada en la inducción.

Gráfica 2



Al utilizar notación matricial y al tratar a los vectores de n componentes como matrices de n renglones y una columna podemos escribir:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

o equivalentemente,

$$d^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y}) \qquad (2)$$

Donde (**x**-**y**)' es el transpuesto de (**x**-**y**). Esta notación nos permitirá visualizar mejor las adaptaciones que se han hecho a la Distancia Euclidiana.

0. Supuestos de la Distancia Euclidiana

Con el fin de descubrir los supuestos inherentes a la Distancia Euclidiana nos referimos al caso de dos dimensiones, esto es, n=2 y consideramos un triángulo rectángulo. El triángulo rectángulo queda definido por los puntos (0,0), (x_1,0) y (x_1,x_2). Los *catetos* son x_1 y x_2 y la *hipotenusa* es la raíz cuadrada de $x_1^2 + x_2^2$. Como los catetos son longitudes, tanto x_1 como x_2 deben ser números reales positivos medidos en ciertas unidades de longitud (centímetros, decímetros, metros, kilómetros, etc.). La restricción de que x_1 y x_2 sean positivos no es necesaria en general para aplicar la fórmula (1).

El Teorema de Pitágoras supone que tanto x_1 como x_2 están medidos en las *mismas* unidades. Esto es, si x_1 está medido en metros entonces x_2 también debe estar medido en metros. Si x_1 estuviera medido en metros y x_2 en pies entonces, $x_1^2 + x_2^2$ combinaría metros al cuadrado con pies al cuadrado por lo que la suma no tiene unidades (estamos sumando peras con manzanas) y el extraer la raíz cuadrada no soluciona el problema, por lo que la hipotenusa tampoco tiene unidades.

Si x_1 y x_2 están medidos en unidades distintas, el Teorema de Pitágoras produce una respuesta incorrecta. Por ejemplo, si x_1=3 metros y x_2=4 pies y utilizamos el Teorema de Pitágoras para calcular la hipotenusa, es claro que la hipotenusa no mide ni 5 metros ni 5 pies. En cambio, si x_1=3 metros y x_2=4 metros, podemos verificar, al medir la hipotenusa con una cinta métrica, que la hipotenusa mide 5 metros.

Al considerar un triángulo rectángulo y graficarlo en un sistema de dos ejes, estamos haciendo el supuesto de que los ejes son perpendiculares (suponemos un sistema *Cartesiano*). En un sistema donde los ejes no son perpendiculares el Teorema de Pitágoras no aplica.

Existen otros supuestos que pueden motivarse a partir de las aplicaciones estadísticas y que posponemos a la siguiente sección de este artículo. Las observaciones hechas anteriormente aplican a cualquier número finito de dimensiones.

1. La Distancia en Aplicaciones Estadísticas

En aplicaciones de la Estadística Multivariada, se caracterizan a personas, marcas, empresas, etc., por medio de vectores de mediciones. Así, una persona

puede estar caracterizada por: estatura (=x_1), edad (=x_2), sexo (=x_3), nivel socioeconómico (=x_4), peso (=x_5), marca de shampoo que usa con mayor frecuencia (x_6), etc. De tal manera que la persona k en una muestra es representada por un vector $\mathbf{x}_k$. Note el lector que las variables pueden ser medidas en escalas diferentes (metros, años, escalas nominales, kilogramos, etc.).

Diversas técnicas estadísticas, como el Análisis de Conglomerados (Cluster Analysis) y el Análisis de Factores (Factor Analysis), están basadas en la *distancia* que existe entre personas o la distancia de las personas a una recta en el espacio. La pregunta que surge de inmediato es: ¿utilizamos la Distancia Euclidiana? En principio, podríamos decir que si las escalas de medición son distintas no deberíamos hacerlo. Pero si perdemos la interpretación de la distancia, ¿es esto razón suficiente para desechar la Distancia Euclidiana? Alguien podría argumentar, "todas las distancias están calculadas de la misma manera y en consecuencia no deberíamos preocuparnos". El ejemplo siguiente nos dice que esto puede no ser prudente.

En la Tabla 1 mostramos el Peso y la Estatura de 17 personas. Esto es, caracterizamos a una persona por medio de un vector de dos componentes (x_1=peso, x_2=estatura) donde x_1 está medida en kilogramos y x_2 en metros.

Tabla 1

Persona	x1 = Peso	x2 = Estatura
1	45	1.52
2	50	1.57
3	54	1.63
4	58	1.60
5	60	1.75
6	62	1.75
7	62	1.53
8	63	1.65
9	65	1.67
10	67	1.74
11	70	1.79
12	72	1.69
13	72	1.65
14	74	1.73
15	76	1.72
16	82	1.87
17	85	1.85
Media	65.7	1.689
Varianza	115.7	0.010
Desv. Est.	10.8	0.101
Correlación	0.794	

Si calculamos el cuadrado de la Distancia Euclidiana entre las personas 5 y 9 tenemos que:

$$d^2(\mathbf{x}_5,\mathbf{x}_9) = (60-65)^2 + (1.75-1.67)^2 = 25+0.0064 = 25.0064$$

Y nos damos cuenta que la diferencia en el peso es la determinante en el cálculo de la distancia. Lo mismo sucede con otros pares de personas; el peso es el factor crucial mientras que la estatura casi no afecta. Pero este hecho es derivado de la escala de medición; una unidad de medición en el peso es un kilo mientras que una unidad de medición en la estatura es un metro. Para que una persona (x_1,x_2) del mismo peso que la persona 5 (x_1=60) esté a la misma distancia que la persona 9 (5) se necesita que la persona tuviera una estatura de 6.75 metros aproximadamente (que se obtiene al resolver $25.0064 = (60-60)^2 + (x_2-1.75)^2$). Note el lector que la variable Peso tiene un varianza mayor que la variable Estatura. La variable de mayor varianza dominará en el cálculo de las distancias (en el Apéndice el lector interesado encontrará la razón).

Si expresamos la Estatura en centímetros (esto es, si hacemos un cambio de escala) entonces:

$$d^2(\mathbf{x}_5,\mathbf{x}_9) = (60-65)^2 + (175-167)^2 = 25+64 = 89$$

Y resulta que la Estatura, para este par de personas, es la determinante en el cálculo de la distancia. Observe el lector lo que produjo el cambio de escala; una unidad en la Estatura es ahora un centímetro; antes una unidad era un metro.

En consecuencia, la escala de medición tiene un impacto directo sobre la distancia. De esta manera, al cambiar la escala de una variable podemos hacer que pese más o menos en el cálculo de la distancia; al utilizar una escala más *fina* incrementamos su peso y al utilizar una escala más *gruesa* le restamos peso. Imagine el lector lo que sucedería si expresáramos

en el ejemplo anterior la Estatura en milímetros y el Peso en toneladas. Una pregunta que surge es, ¿cuál es la escala óptima para cada variable? E inmediatamente nos preguntamos, ¿óptima en qué sentido? El que esto escribe no ha encontrado las respuestas. Anderson (1958) en su famoso libro de Estadística Multivariada escribe en la página 278: "El análisis en componentes principales es más adecuado cuando todas las componentes de **X** están medidas en las mismas unidades. Si no están medidas en las mismas unidades, el racional de maximizar $\beta'\Sigma\beta$ relativo a $\beta'\beta$ es cuestionable; de hecho, el análisis dependerá de las diversas unidades de medición". Desdichadamente, no nos dice cuál es la escala más adecuada.

Al expresar una variable en una escala más *fina* hacemos que la desviación estándar (o equivalentemente la varianza) sea mayor. En consecuencia, el problema de las escalas distintas está relacionado con la situación en la que las distintas variables tienen varianzas distintas.

Para resolver el problema de las distintas escalas de medición se han hecho varias propuestas que pueden resumirse como: escale los datos respecto a una *estadística* (use índices) y luego aplique la distancia Euclidiana. Por ejemplo, si la estadística es la *media aritmética* transformaríamos los datos a índices respecto a las distintas medias:

$$\mathbf{x}=(x_1, x_2,\ldots,x_n)' \rightarrow \mathbf{x}=(x_1/m_1, x_2/m_2, \ldots, x_n/m_n)'$$

Donde las m's son las medias aritméticas. El cuadrado de la distancia entre dos vectores **x** y **y** estaría dado por la fórmula:

$$(x_1-y_1)^2/m_1^2 + (x_2-y_2)^2/m_2^2 + \ldots + (x_n-y_n)^2/m_n^2$$

El lector puede darse cuenta que se pueden utilizar índices respecto a mínimos, máximos, medianas, etc.

Existe una propuesta que sugiere escalar los datos por las *desviaciones estándar*. Esta propuesta logra dos objetivos; por una parte elimina la escala y por otro logra que todas las variables tengan la misma varianza (igual a uno). Así, el cuadrado de la distancia entre dos vectores **x** y **y** estaría dado por la fórmula:

$$(x_1-y_1)^2/S_1^2 + (x_2-y_2)^2/S_2^2 + \ldots + (x_n-y_n)^2/S_n^2$$

Donde las S's son las desviaciones estándar. Note el lector que al escalar por las desviaciones estándar las variables quedan expresadas en unidades que no necesariamente son fáciles de interpretar.

Los escalamientos planteados llevan a la siguiente variación en la fórmula (1):

$$d_1(\mathbf{x},\mathbf{y})=\sqrt{w_1(x_1-y_1)^2 + w_2(x_2-y_2)^2+\ldots+w_n(x_n-y_n)^2}$$

Donde $w_1, w_2, \ldots, w_n$ son los factores de escalamiento.

Por otra parte las w's pueden ser interpretadas como *pesos* o *importancias* de cada variable o dimensión y son números reales no negativos. Mientras una w sea mayor, mayor será la influencia de esa variable en el cálculo de la distancia. Si todos los pesos son iguales, esto es, $w_1=w_2=\ldots=w_n$, estaremos aplicando esencialmente la Distancia Euclidiana (en consecuencia, la Distancia Euclidiana da el mismo peso a todas las variables).

Si reconsideramos el ejemplo presentado arriba (Peso en kilos y Estatura en centímetros), pero asignamos pesos distintos a las variables, por ejemplo: $w_1=4$ y $w_2=1$, (las comisiones de Box de todo el mundo enfrentan peleadores por peso no por estatura) tendremos que:

$$d_1^2(\mathbf{x}_5,\mathbf{x}_9) = 4(60-65)^2 + (175-167)^2 = 100+64 = 164$$

Y el peso vuelve a ser el factor determinante en el cálculo de la distancia. Pero surge la pregunta ¿por qué asignar al Peso 4 veces más importancia que a la Estatura?, ¿por qué 4 y no 10 u otro número?, ¿bajo qué criterio de optimalidad podemos seleccionar a las w's? La pregunta no está respondida en los libros de Estadística Multivariada.

Si definimos la matriz **W** como aquella matriz que tiene en la diagonal principal a las w's y fuera de la diagonal ceros, esto es, $\mathbf{W}=\text{diag}(w_1, w_2, \ldots, w_n)$ podemos escribir:

$$d_1^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})'\mathbf{W}(\mathbf{x}-\mathbf{y}) \qquad (2)$$

Y es posible demostrar que d_1 es una distancia (satisface las tres condiciones mencionadas en la sección 1 del artículo).

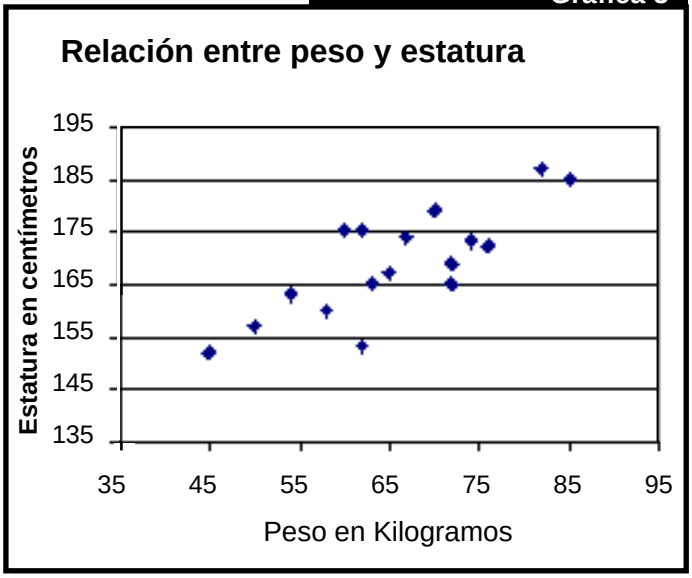
Uno de los análisis multivariados que permite motivar una nueva modificación a la fórmula de la Distancia Euclidiana es la detección de *observaciones aberrantes* (outliers en inglés). Una observación aberrante es un vector de datos que es *muy diferente* del resto. Estas observaciones pueden distorsionar seriamen-

te el análisis y deben ser identificadas lo antes posible. La decisión de incluirlas o excluirlas del análisis depende de cada caso particular.

La detección está basada principalmente en el concepto de distancia al vector de medias; si una observación está *muy lejos* del vector de medias es candidata a ser aberrante.

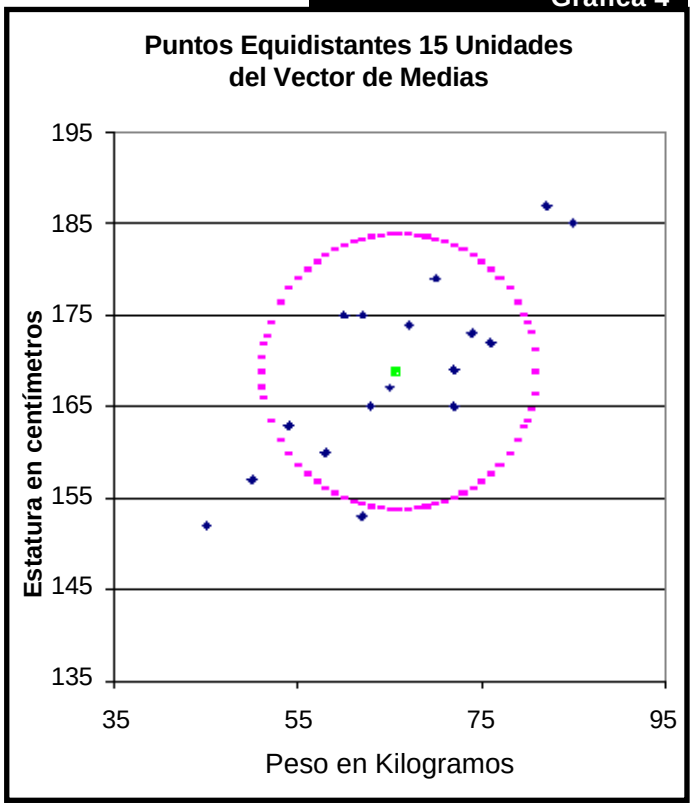
Bajo la Distancia Euclidiana los puntos equidistantes del vector de medias son una hiperesfera (generalización de círculo y esfera) reflejando el hecho de iguales escalas, varianzas e importancias. Bajo la variación presentada de la Distancia Euclidiana, los puntos equidistantes del vector de medias son un hiperelipsoide paralelo a los ejes (generalización de elipse y elipsoide) reflejando la existencia de escalas, varianzas e importancias distintas. Ninguna de las dos distancias toma en cuenta la *correlación* existente entre las variables; vea la Gráfica 3. El Coeficiente de Correlación entre el Peso y la Estatura para las 17 personas es 0.79, el cual es significativamente mayor que cero.

Gráfica 3



Para ilustrar la importancia de la correlación, supongamos que decidimos tentativamente considerar como aberrantes aquellas observaciones que distan 15 *unidades* o más del vector de medias según la Distancia Euclidiana. En la Gráfica 4 mostramos los datos originales y el círculo de radio 15 centrado en el vector de medias. De acuerdo al criterio tentativo, las personas 1,2,7,16 y 17 son candidatas a ser aberrantes. Note el lector que las desviaciones estándar son muy parecidas (aproximadamente 10).

Gráfica 4

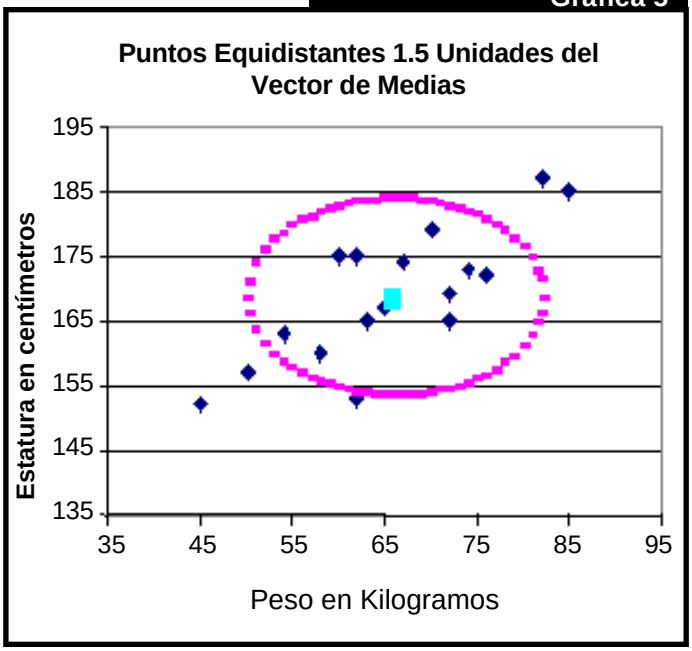


Si el criterio tentativo para detectar observaciones aberrantes es el que una observación diste más de 1.5 *unidades* del vector de medias pero en la Distancia que escala por medio de la desviación estándar, entonces las observaciones candidatas a ser aberrantes son 1,2,7,16 y 17 (las mismas que con la Distancia Euclidiana; esto es debido a que las desviaciones estándar son muy similares). La Gráfica 5 muestra los datos originales y la elipse que permite hacer la detección.

Con ambas distancias, Euclidiana y su variación, una persona de 80 kilos de Peso y con Estatura de 180 centímetros sería candidata a ser aberrante, pero ambos considerarían *normales* a las personas (58,180) y (75,158). Es claro que personas con esas últimas características parecerían pertenecer a un grupo distinto del de los 17 listados, ya que rompen con el patrón observado. Ese patrón está reflejado por la correlación existente entre las variables Peso y Estatura.

La Distancia de Mahalanobis toma en cuenta la correlación existente entre las variables estudiadas para determinar la distancia. Si la matriz de varianzas y covarianzas de las variables estudiadas es **S**, el cua-

Gráfica 5



drado de la distancia de Mahalanobis entre los vectores **x** y **y** está definida como:

$$d_M^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})$$

Donde $\mathbf{S}^{-1}$ es la matriz inversa de **S**. Con el fin de ver cómo interviene la correlación entre variables revisaremos el caso particular n=2. Cuando dos variables intervienen en el análisis, la matriz **S** tiene en la diagonal las varianzas de las variables y fuera de la diagonal tiene la covarianza, esto es:

$$\mathbf{S} = \begin{pmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{pmatrix}$$

De manera explícita tenemos que:

$$d_M^2(\mathbf{x},\mathbf{y}) = [(x_1-y_1)^2/S_1^2 + (x_2-y_2)^2/S_2^2 - 2r(x_1-y_1)(x_2-y_2)/(S_1S_2)]/(1-r^2)$$

Donde r es el coeficiente de correlación entre x_1 y x_2, definido como la covarianza divida entre el producto de las desviaciones estándar, esto es:

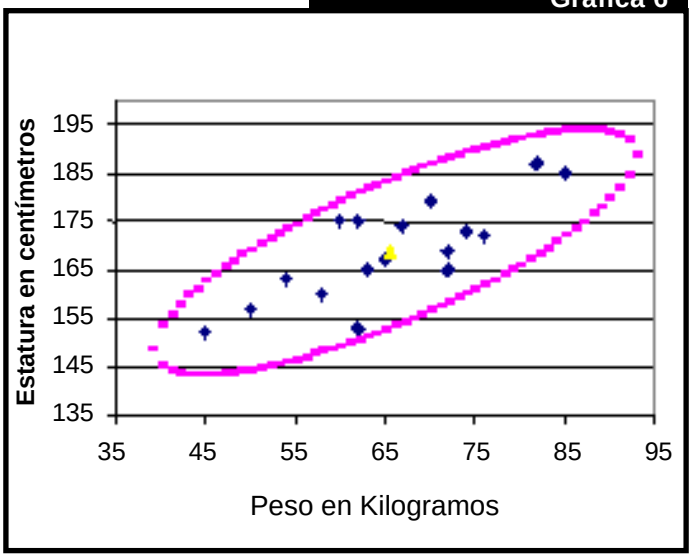
$$r = S_{12}/(S_1S_2)$$

Observemos primero que si r=0, la Distancia de Mahalanobis se reduce a la variación de la Distancia Euclidiana que hace un escalamiento por la desviación estándar. Cuando r≠0, los dos primeros términos corresponden a la Distancia Euclidiana con las variables escaladas por la desviación estándar, pero afectadas por el término $1/(1-r^2)$. El tercer término es la novedad.

En la Gráfica 6 mostramos la elipse que define los puntos candidatos a ser considerados aberrantes. Note el lector que la elipse no tiene sus ejes paralelos a los ejes del sistema de dos coordenadas. Este hecho se debe a la correlación existente entre el Peso y la Estatura. Notamos que ninguna de las observaciones es candidata a ser aberrante. Por otra parte, personas con vectores de mediciones (58,180) y (75,158) serían consideradas tentativamente como aberrantes (que no sucedería con las otras dos distancias consideradas previamente) a pesar de estar *más cerca*, desde el punto de vista de la Distancia Euclidiana, del vector de medias que la observación (85,185).

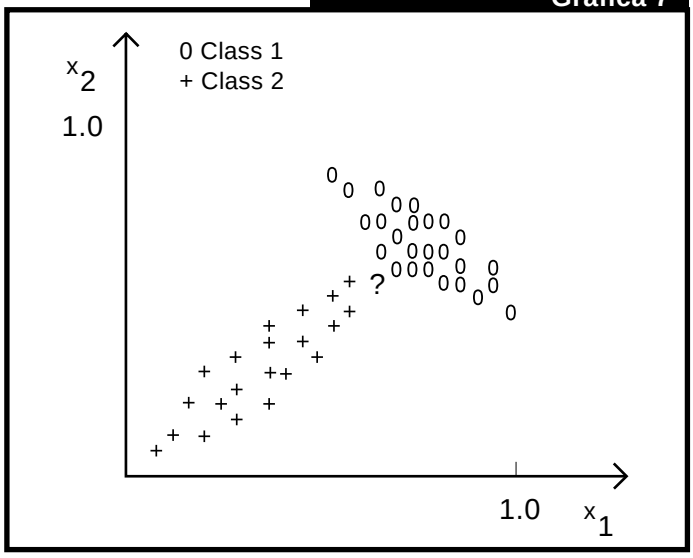
Parecería que la distancia de Mahalanobis incorpora una tercera dimensión en donde el centro de la elipse es la cumbre de una colina. Las elipses que representan los puntos equidistantes, serían curvas de nivel de la colina. En la dirección de la correlación la colina tiene una pendiente suave mientras que en la dirección perpendicular la colina es escarpada con una pendiente grande. Esto explicaría el por qué datos que aparentemente están cerca en realidad están lejos.

Gráfica 6



Otro ejemplo de la aplicación de la distancia de Mahalanobis es en la clasificación de observaciones en grupos o poblaciones. El siguiente ejemplo está tomado de "9.6.2 Distance Measures in a Feature Space", (http://www.maths.uwa.edu.au/~rkealley/ann_all/node125.html). Vea la Gráfica 7 de la siguiente página.

Gráfica 7



En la gráfica se muestran observaciones de dos Distribuciones Normales bivariadas (identificadas en la gráfica como Class 1 y Class 2). La gráfica también muestra una observación marcada "?" y surge la pregunta ¿a cuál de las dos Distribuciones asignamos "?" ? La distancia Euclidiana sugiere que "?" pertenece a Class 1. Pero si aplicamos la Distancia de Mahalanobis, "?" será asignada a Class 2. La razón de esta diferencia es que la Distancia de Mahalanobis toma en cuenta tanto la diferencia en desviaciones estándar como la correlación existente, mientras que la Distancia Euclidiana no lo hace.

La distancia de Mahalanobis nos permite ver que la Distancia Euclidiana tiene implícito el supuesto de que las variables involucradas no están correlacionadas.

La distancia de Mahalonobis es mencionada en diversos libros de Estadística Multivariada; por ejemplo, Anderson (1958), Morrison (1976), Hair, Anderson, Tatham y Black (1995) y Grimm y Yarnold (1998), pero ninguno nos dice que es óptima en algún sentido. Distintos paquetes de computación tienen como opción de distancia la Distancia de Mahalanobis para realizar los cálculos.

De los ejemplos presentados concluimos que la Distancia de Mahalanobis resulta más apropiada en las aplicaciones de la Estadística Multivariada que la Distancia Euclidiana, ya que se toma en cuenta no nada más el hecho de que las variables estén medidas en distintas unidades, sino también toma en cuenta la posibilidad de varianzas distintas y la correlación entre variables.

Desde un punto de vista estrictamente matemático, las distancias que hemos considerado aquí tienen la forma general siguiente:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\mathbf{B}(\mathbf{x}-\mathbf{y})}$$

Donde **x** y **y** son vectores de n coordenadas y **B** es una matriz nxn que es simétrica y positiva definida.

Si **B** es la matriz idéntica, obtenemos la Distancia Euclidiana; si **B** es una matriz diagonal con los inversos multiplicativos de las varianzas, obtenemos la Distancia que escala por medio de las desviaciones estándar; y si **B** es la inversa de una matriz de varianzas y covarianzas, obtenemos la Distancia de Mahalanobis.

Los supuestos hechos sobre la matriz **B**, nos permiten establecer (según el Teorema de la Descomposición Espectral) que existen dos matrices **V** y Λ, donde **V** es ortogonal (**VV'=I= V'V)** y Λ es diagonal tales que, **B**=**V**Λ**V**' y en consecuencia:

$$d^2(\mathbf{x},\mathbf{y}) = [\mathbf{V'(x-y)}]' \Lambda [\mathbf{V'(x-y)}]$$

Lo interesante de la fórmula es que **V'(x-y)** es el vector de coordenadas de (**x**-**y**) según la base definida por las columnas de **V** y la distancia es del tipo de la dada en la fórmula (2), donde los pesos de las distintas dimensiones están establecidos por los valores característicos de **B**, esto es, por los elementos de $\Lambda=\text{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n)$. De esta manera, la ponderación de las distintas dimensiones es determinada por un procedimiento que no depende del criterio del que analiza los datos.

Los valores característicos de **B**, λ_1, λ_2,…, λ_n, están asociados a diversas características de la matriz **B** en particular, *el determinante* de **B** es el producto de los valores característicos (de tal manera que la matriz es singular siempre y cuando al menos un valor característico valga cero) y *la traza* de **B** vale la suma de los valores característicos.

Apéndice

La razón del por qué la variable con mayor varianza dominará las distancias es la siguiente: suponga que **X** es un vector aleatorio de n componentes que tiene media μ (=E(**X**)) y matriz de varianzas y covarianzas Σ(=E(**X**-μ)(**X**-μ)'). Al calcular el promedio del cuadrado de la distancia de **X** a μ tenemos que:

$$E[d^2(\mathbf{X},\mu)]=E[(\mathbf{X}-\mu)'(\mathbf{X}-\mu)]=E[\text{tr}(\mathbf{X}-\mu)(\mathbf{X}-\mu)']=\text{tr}(\Sigma)=\sum_{k=1}^{n}\text{Var}(X_k)$$

Donde tr(•) denota la traza de una matriz. La fórmula permite ver que la variable con la mayor varianza será la que domine la suma (que es el cuadrado de la distancia promedio a la media).

Los cálculos numéricos en relación con la detección de observaciones aberrantes fueron hechos al determinar los puntos equidistantes a 15 unidades, 1.5 unidades y 2.5 unidades del vector de medias. La razón de utilizar esos números y hacerlos comparables es la siguiente: para la Distancia Euclidiana 15 unidades representan aproximadamente 1.5 veces la desviación estándar común. Para la distancia que escala por las desviaciones estándar, 1.5 unidades representan 1.5 veces la desviación estándar común, aproximadamente 10 (en total 15 unidades). Para la distancia de Mahalanobis, 2.5 unidades representan 2.5 veces la desviación estándar común multiplicada por la raíz cuadrada de uno menos el cuadrado del coeficiente de correlación (aproximadamente 15 unidades en total).

Una matriz **V** es ortogonal cuando sus columnas forman una base ortonormal, este hecho es consecuencia de la igualdad **V**'**V**=**I**. A la vez, **V**' resulta ser la inversa de **V**.

Las columnas de una matriz **B** definen un paralelotopo (generalización de paralelogramo y paralelepípedo), y el volumen de dicho paralelotopo es igual al valor absoluto del determinante de la matriz. En consecuencia, que el determinante de una matriz valga cero es equivalente a que el paralelotopo tenga volumen cero. Cuando la matriz tiene 2 columnas, el paralelogramo tiene área cero cuando las dos columnas son paralelas, esto es, cuando *falta* una dimensión.

Bibliografía

Anderson, T.W. (1958), *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, New York, New York.

Grimm, L.G. and Yarnold, P.R. Editors (1998), *Reading and Understanding Multivariate Statistics*, American Psychological Association, Washington, D.C.

Hair, J.F., Anderson, R.E., Tatham, R.L. y Black, W.C. (1995), *Multivariate Data Analysis*, Fourth Edition, Prentice Hall, Upper Saddle River, New Jersey.

Kolmogorov, A.N. y Fomin, S.V. (1970), *Introductory Real Analysis*, Prentice Hall, Inc., Englewood Cliffs, N.J.

Morrison, D.F. (1976), *Multivariate Statistical Methods, Second Edition*, McGraw-Hill Book Company, New York, New York.