matemáticas: Una publicación del ICM – ESPOL
2009, Vol. 7, No. 2

EFECTO DE LA DIMENSIÓN, LA DISTANCIA ENTRE LOS CENTROIDES Y EL SESGO EN UNA VARIABLE, EN DISCRIMINACIÓN LINEAL PARA DOS GRUPOS

Serna César[1]

Resumen. *En este documento se discute el método de Análisis Lineal Discriminante (LDA) y la Distribución Normal Sesgada (SN) (Azzalini & Dalla Valle), dos técnicas estadísticas que combinadas sirven para diagnosticar comportamientos reales de poblaciones que se suponen tienen un comportamiento aproximadamente normal. Para ello se generan poblaciones normales multivariadas puras y poblaciones normales multivariadas que poseen una variable con sesgo positivo. En este trabajo se determinan las tasas de error de mala clasificación de las poblaciones normales multivariadas puras y se compara con la tasa de error de mala clasificación de las poblaciones normales que poseen la variable sesgada. Estas tasas de mala clasificación se determinan, haciendo variar la dimensión de estudio (p=2,.. 20); la distancia Mahalanobis entre los centroides de las dos poblaciones de estudio, y el grado del sesgo w; cuando se generan 1000 individuos o unidades experimentales de cada población. Además se determinan estas situaciones desde el punto de vista muestral suponiendo que las poblaciones pseudogeneradas son muestras de tamaño 1000; esto con el fin de determinar diferencias entre lo poblacional y lo muestral. También se generan múltiples gráficas que sirven para visualizar tanto el comportamiento poblacional como el muestral de las tasas de error de mala clasificación. Es de interés señalar que en el escenario muestral, a medida que aumenta la dimensión de estudio disminuye la tasa de error de mala clasificación, mientras que en el caso poblacional esta situación es mucho más débil de percibir.*

Palabras Clave: Normal Sesgada, APER, Discriminación, Distancia Mahalanobis, Sesgo.

1. INTRODUCCIÓN

La Función Lineal Discriminante (LDF) es comúnmente usada en la práctica debido a que dentro de las posibilidades que ofrece la mayoría de los programas de cómputo la incluyen como opción para llevar a cabo análisis discriminante e igualmente ha sido estudiada y discutida por muchos autores desde diferentes puntos de vista para evaluar su desempeño discriminatorio. Como es evidente, la información que los paquetes procesan es por lo general información obtenida a partir de muestras; estos paquetes utilizan reglas de clasificación que se basan en estimaciones de los parámetros asociados a la regla derivada de aspectos a optimizar como lo es comúnmente el costo esperado de mala clasificación. Justamente el énfasis de su estudio se ha orientado hacia la calidad y robustez de las estimaciones y el efecto de los tamaños de muestra, al compararla con otras reglas de clasificación; y como en su esencia reside el supuesto de homoscedasticidad (Wahl y Kronmal, 1977), esto ha motivado otro tema central en el estudio de la misma. El objeto de este trabajo no está propiamente en continuar dentro del estudio de situaciones de robustez en la estimación de los parámetros de la LDF, ni en el efecto de asumir homoscedasticidad cuando no hay razones que la justifiquen. El objetivo es detectar el efecto que tiene; la *lejanía* o *cercanía* de las poblaciones de estudio y del número de variables estudiadas ante la presencia de una variable con un comportamiento singular o atípico respecto al comportamiento de las otras variables elegidas, en las tasas de clasificación incorrecta, para efectos de la misma discriminación. Para ello se realiza simulaciones como estrategia metodológica para acopiar evidencias que permitan evaluar la importancia del efecto señalado. Los datos sintéticos producidos por un proceso de simulación son útiles en la descripción del desempeño de un procedimiento estadístico que permiten elegirlo o sustituirlo por procedimientos alternativos en casos puntuales. Discriminar en estadística es una actividad muy particular que conlleva riesgos bien conocidos, por tanto el tener elementos de juicio del desempeño de una regla de clasificación permite elegir entre diferentes opciones para llevar a cabo la tarea discriminatoria.

2. MARCO TEÓRICO

El objetivo del análisis discriminante es localizar un nuevo individuo en una de estas poblaciones, con base en la información contenida en x_0. Es claro que esta metodología, aunque no es lo deseable, produce clasificaciones incorrectas lo cual se señala más adelante. Usualmente no se tiene información a priori de la población de donde probablemente proviene cada individuo; sin embargo, si tal información está disponible, se podrá incorporar dentro de una aproximación bayesiana. La situación donde la función de distribución de probabilidad es exactamente conocida, es básicamente un análisis teórico de la situación. Una variante de esta situación ocurre cuando la forma de la función de distribución de probabilidad para cada población es desconocida, pues en este caso los parámetros son estimados a través de muestras (Mardia, Kent y Bibby; 1992).

[1] *Serna Mejía César Augusto, MSc. Estadística. Universidad Nacional de Colombia. Docente de Estadística. Universidad Central. Bogotá – Colombia. (e_mail:cesarserna29@gmail.com)*

Análisis Lineal Discriminante para dos Grupos:

Sean $\mathbf{f_1(X)}$ y $\mathbf{f_2(X)}$ las funciones de densidad asociadas al vector de variables aleatorias **X** para las poblaciones π_1 y π_2, respectivamente. Un objeto o individuo con medidas asociadas **x** debe ser asignado en una de las dos poblaciones. Sea Ω el espacio muestral, el cual es la colección de todas las posibles observaciones **x**. Sea $\mathbf{R_1}$ el conjunto de valores de **x** para los cuales se localiza el objeto o individuo en π_1 y sea $\mathbf{R_2} = \Omega - \mathbf{R_1}$ los restantes valores de **x**, para los cuales se localiza el objeto o individuo en π_2. Como todos los individuos deben ser asignados a una y solo una población, los conjuntos $\mathbf{R_1}$ y $\mathbf{R_2}$ son mutuamente excluyentes y exhaustivos (R. A. Johnson y Wichern, 2002). La probabilidad condicional P(2|1) de clasificar un objeto como π_2 cuando, en efecto, éste pertenece a π_1, es $P(2 \mid 1) = P\left(X \in R_2 \mid \pi_1\right) = \int_{R_2} f_1(x)dx$ y similarmente la probabilidad condicional **P(2|1)** de clasificar un objeto como π_1 cuando en realidad pertenece a π_2, es $P(1 \mid 2) = P(X \in R_1 \mid \pi_2) = \int_{R_1} f_2(x)dx$.

Sea p_1 la probabilidad a priori de π_1 y p_2 la probabilidad a priori de π_2; donde se cumple que $p_1 + p_2 = 1$. Las probabilidades de asignar correcta o incorrectamente un individuo, pueden ser derivadas como el producto de las probabilidades condicionales y las a priori, así:

P(asignar correctamente un objeto como π_1) = $P(X \in R_1 \mid \pi_1).P(\pi_1) = P(1 \mid 1).p_1$

P(asignar correctamente un objeto como π_1) = $P(X \in R_1 \mid \pi_2).P(\pi_2) = P(1 \mid 2).p_2$

P(asignar correctamente un objeto como π_2) = $P(X \in R_2 \mid \pi_2).P(\pi_2) = P(2 \mid 2).p_2$

P(asignar correctamente un objeto como π_2) = $P(X \in R_2 \mid \pi_1).P(\pi_1) = P(2 \mid 1).p_1$

Las reglas de discriminación tienen un componente adicional llamado costo de clasificación incorrecta $C(i \mid j)$ el cual implica pagar un precio (no solo económico) por clasificar incorrectamente un conjunto de individuos. Cuando las probabilidades a priori y los costos de mala clasificación son indeterminados, entonces comúnmente se considera cada cociente igual a 1. Procedimientos de clasificación basados en poblaciones normales predominan en la práctica estadística por su simplicidad y gran eficiencia. Para ello se asume que $\mathbf{f_1(x)}$ y $\mathbf{f_2(x)}$ son densidades normales multivariadas; la primera con vector de medias μ_1 y matriz de covarianza Σ_1 y la segunda con vector de medias μ_2 y matriz de covarianza Σ_2. El caso especial de igual matriz de covarianzas ($\Sigma 1 = \Sigma_2 = \Sigma$) conduce a una particular estadística de clasificación llamada función lineal discriminante (LDF). Suponga que la densidad conjunta de $X' = (X_1, X_2, ..., X_p)$; para las poblaciones π_1 y π_2 son dadas por

$$f_i = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{p}{2}}} . \exp\left[-\frac{1}{2}(x - \mu_i)' \Sigma^{-1} (x - \mu_i) \right]$$

con i=1,2. Suponga también que los parámetros poblacionales 1 y 2 son conocidos. Entonces, después de eliminar las constantes y simplificar

$$R_1 = \exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2) \right] \geq \left[\frac{c(1|2)}{c(2|1)} \right] \left(\frac{p_2}{p_1} \right)$$

$$R_2 = \exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2) \right] < \left[\frac{c(1|2)}{c(2|1)} \right] \left(\frac{p_2}{p_1} \right)$$

la expresión, se tiene que las regiones que minimizan el error de clasificación incorrecta, son de la forma:

Dadas las regiones $\mathbf{R_1}$ y $\mathbf{R_2}$; podemos construir la

$$(\mu_1 - \mu_2)'\Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq \ln\left[\left[\frac{c(1|2)}{c(2|1)} \right] \left(\frac{p_2}{p_1} \right) \right]$$

regla de clasificación que minimiza el ECM, como sigue. Localizar de $\mathbf{x_0}$, en la población π_1, si:

Localizar de $\mathbf{x_0}$, en la población π_2, en caso contrario. Cuando no se conoce los costos de mala clasificación ni las probabilidades a priori, generalmente se asumen proporcionales, luego estos cocientes se reemplazan por 1 y por tanto la regla de clasificación es de la forma:

Localizar de $\mathbf{x_0}$, en la población π_1, si:

$$(\mu_1 - \mu_2)'\Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq 0$$

Localizar de $\mathbf{x_0}$, en la población π_2, en caso contrario.

Una forma más eficiente para medir el desempeño de la regla y por consiguiente calcular estas tasas de error de mala clasificación, es por medio de la Tasa de Error Aparente APER (Apparent Error Rate). El APER, puede ser fácilmente calculado a través de la matriz confusión, de la siguiente manera:

		Membresías Predichas	
		π_1	π_2
Membresías Actuales	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}

Donde:

n_{1C} = Número de ítems de π_1 correctamente clasificados como π_1

n_{2M} =Número de ítems de π_2 incorrectamente clasificados como π_1

La tasa de error aparente es entonces $APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$; que es la proporción de ítems en la muestra que son mal clasificados (R. A. Johnson y Wichern, 2002). Una buena regla de clasificación debe tener una APER, lo más pequeña posible.

Distribución Normal Multivariada Sesgada:

El término *Normal Sesgada* (*SN*) hace referencia inicialmente a una clase paramétrica de distribuciones probabilísticas que incluyen la distribución normal estándar como caso particular. Una variable aleatoria Z puede ser normal sesgada, con parámetro de sesgo λ, y se nota como $Z \sim SN(\lambda)$, si su función de densidad es de la forma $\phi(Z;\lambda) = 2\phi(Z) * \Phi(\lambda Z);\ Z \in \mathbb{R}$; donde $\phi(Z)$ y $\Phi(\lambda Z)$ denota la función de densidad y la función de distribución de una normal estándar, respectivamente. El parámetro λ que regula el sesgo varía entre $(-\infty;\infty)$, y si $\lambda = 0$; la distribución resultante corresponde precisamente a la de una $N(0;1)$ (Azzalini y Dalla Valle; 1996). Desde el punto de vista aplicativo, la densidad anterior es apropiada para simular el comportamiento de poblaciones que muestran distribuciones empíricas unimodales con determinada presencia de sesgo, la cual es una situación que a menudo ocurre en los problemas prácticos. La función generadora de momentos y los primeros momentos de Z son dados por Azzalini (1988); en particular se tiene que $E[Z] = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \delta$ y $Var[Z] = 1 - \frac{1}{2}\delta^2$ donde δ está relacionada con λ a través de la siguiente expresión: $\delta(\lambda) = \frac{\lambda}{\left(1+\lambda^2\right)^{\frac{1}{2}}}$ y $\lambda(\delta) = \frac{\delta}{\left(1-\delta^2\right)^{\frac{1}{2}}}$. El índice del sesgo de este método varía en el intervalo (-0.995; 0.995). Se tiene que si Y y W son variables independientes $N(0;1)$, y Z es un conjunto igual a Y condicionado en Y > W; para algún λ real, entonces $Z \sim SN(\lambda)$. Este resultado es esencialmente equivalente a probar que $\phi(Z;\lambda)$ es una función de densidad dada en Azzalini (1985). Para generar números aleatorios, es más eficiente usar una variante de este resultado; a saber $Z = \begin{cases} Y \text{ si } \lambda Y > W \\ -Y \text{ si } \lambda Y \le W \end{cases}$.

El caso multivariado del método propuesto por Azzalini y Dalla Valle, es tomado para resaltar algunas propiedades de la distribución *SN*. Para expresar la extensión multivariada k-dimensional de $\phi(Z;\lambda)$; es necesario considerar una variable **Z** multivariada tal que cada una de sus componentes es normal sesgada. Por tanto se define la distribución conjunta de **Z** como ***SN*** multivariada. Considere una variable aleatoria normal k-dimensional $Y = (Y_1, Y_2, ..., Y_K)'$ con marginales estandarizadas, independientes de $Y_0 \sim N(0;1)$; así $\begin{pmatrix} Y_0 \\ \boldsymbol{Y} \end{pmatrix} \sim \boldsymbol{N}_{k+1}\left\{0, \begin{pmatrix} 1 & 0 \\ 0 & \psi \end{pmatrix}\right\}$ donde ψ es una matriz de orden KxK que representa la correlación de **Y**. Sean $(\delta_1, \delta_2, ..., \delta_k)$ valores entre (−1;1); entonces se definen las distribuciones normales estándar marginales como $Z_j = \delta_j |Y_0| + \left(1-\delta_j^2\right)^{\frac{1}{2}}$ con $j = 1,2,...,k$; tal que $Z_j \sim SN\{\lambda(\delta)\}$. El cálculo de la distribución de $Z = (Z_1, Z_2, ..., Z_K)'$; es fácil pero largo, la expresión sintetizada es $f_k = 2\phi_k(Z,\Omega).\Phi(\alpha' Z);\ Z \in \mathbb{R}$. Se puede afirmar que una variable aleatoria **Z** con función de densidad $f_k(Z)$ es una variable normal sesgada k-dimensional, con vector de parámetros de forma λ y parámetro de dependencia ψ; para abreviar se puede escribir $Z \sim SN_k(\lambda;\psi)$. Si Z tiene función de densidad $f_k(Z)$; entonces su función de distribución acumulada, donde la variable aleatoria $Z \in \mathbb{R}^k$, es como se define a continuación:

$$F_k = F_k(Z_1, Z_2, ..., Z_K) = 2\int_{-\infty}^{Z_1} ... \int_{-\infty}^{Z_k} \int_{-\infty}^{\alpha'\gamma} \phi(\gamma;\Omega) du d\gamma_1 d\gamma_2 ... d\gamma_k$$

Para concluir, la función de distribución de la variable k-dimensional $Z \sim SN_k(\lambda;\psi)$ puede ser obtenida de F_k, al calcular la función de distribución de una variable normal (k+1)-dimensional con distribución mostrada anteriormente.

3. IMPLEMENTACIÓN COMPUTACIONAL

El proceso de simulación puede entenderse como un proceso experimental que investiga efectos de niveles de factores de interés en el desempeño de un procedimiento estadístico.

Particularmente la discriminación es una herramienta estadística de una utilidad innegable, y como su objeto final es la clasificación, en uno de varios grupos o poblaciones, de unidades estadísticas que presentan información particular, es evidente entonces que la evaluación de la habilidad discriminatoria de un procedimiento en este campo radique en la cantidad de aciertos que el procedimiento evidencie en unidades estadísticas de pertenencia poblacional segura. Por medio de la tasa de error aparente de mala clasificación como respuesta a la modificación de los diferentes niveles de los factores controlados cuya combinación representan diferentes escenarios o situaciones planteadas en este trabajo, se evalúa la habilidad de la regla lineal discriminante frente un cambio sutil en la estructura de los supuestos; presencia de sesgo en una de sus variables. Estos factores como ya se ha mencionado son; sesgo, dimensión y distancia entre centroides. Los 19 niveles elegidos para la dimensión fueron p =2,..,20; para el sesgo w se eligieron 19 niveles desde 0.1 hasta 0.9 con incremento de 0.1; y desde 1.0 hasta 10 con incremento de una unidad. El último factor contó 10 distancias igualmente espaciadas entre 0.5 y 5.0 con incrementos de 0.5. Para cada combinación de niveles, un valor de sesgo, un número determinado de variables y una distancia específica entre centroides, el proceso de cómputo calcula 5000 tasas de error aparente cuando las observaciones son generadas a partir de distribuciones normales sin sesgo, entendidas como testigo en la experimentación, y 5000 cuando las observaciones son generadas a partir de distribuciones normales en presencia de sesgo. En síntesis el programa calcula 36100000 tasas de error aparente y registra como respuesta 3610 tasas promedio, las cuales responden al objetivo central de este trabajo que es determinar las tasas de clasificación incorrecta en el contexto poblacional; adicionalmente este trabajo presenta en las mismas condiciones un contexto suplementario que se denomina contexto muestral que se desarrolló suponiendo que la información seudo-generada son muestras que provienen de poblaciones normales y normales sesgadas. Para simular los diferentes escenarios que permitieron evaluar el efecto de los factores estudiados, dimensión, distancia y sesgo en el cálculo de la tasa de clasificación incorrecta, se implementaron para tal efecto, rutinas propias del ambiente y lenguaje para computación estadística R. Esta herramienta computacional tiene algunas funciones estadísticas ya definidas; sin embargo la mayoría de las rutinas fundamentales para este trabajo al no estar implementadas, fue necesario programarlas y modificar algunas existentes.

El proceso de simulación contó con cuatro pasos generales como lo describe la figura 1, cuyos objetivos son generar parámetros poblacionales, generar unidades de poblaciones normales, ensamblar la regla de clasificación, calcular las tasas de clasificación incorrecta.

El primer paso fue generar los parámetros poblacionales μ_1, μ_2 y Σ; teniendo en cuenta que antes de generar μ_2 se generó Σ y luego μ_2 se generó condicionado a que μ_1 y μ_2 estén a una distancia Mahalanobis D.

En el segundo paso se construyó una rutina cuya función es generar observaciones de poblaciones normales p-variadas, que dentro de su estructura se apoya en el comando rmvnorm (n; mu; cov) del paquete mvtnorm en R, el cual bajo la distribución normal multivariada con vector de medias μ_1 y matriz de covarianzas μ_2 ; genera n observaciones p-variadas que simulan información de individuos que, para efectos del desarrollo computacional, son llamados P_1. De igual manera las n observaciones de la segunda población son generadas a través de la misma función rmvnorm; pero empleando los parámetros $(\mu_2; \Sigma)$. Estas observaciones son denominadas P_2. La rutina generación proporciona una tercera población p-variada con los mismos parámetros de la segunda población, es decir $(\mu_2; \Sigma)$; pero con la única diferencia de que las observaciones generadas provienen de la mencionada distribución normal p-variada con una de sus variables afectada por un sesgo w positivo determinado.

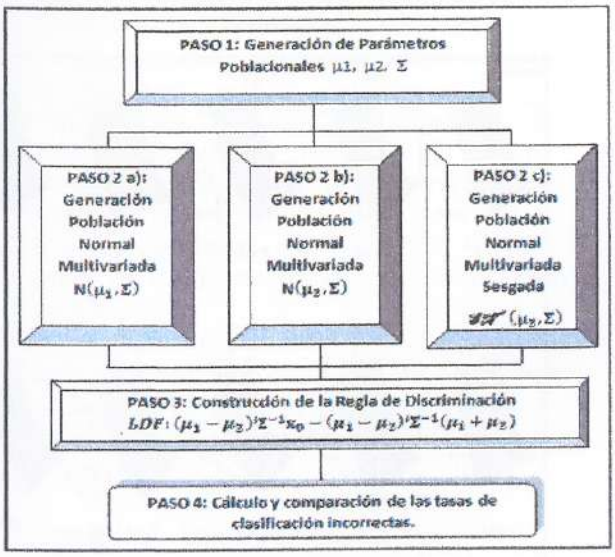
En el tercer paso, aunque es un paso elemental dentro de todo el proceso de simulación y que se reduce a una instrucción simple, se destaca dentro de estos pasos metodológicos para resaltar la esencia del trabajo dirigida a evaluar el comportamiento de la regla de clasificación tanto en su versión poblacional como en su versión muestral, dependiendo como ya se ha mencionado, de los factores distancia, dimensión y sesgo.

El cuarto paso es evaluar la regla a través del proceso de simulación, la forma de llevar a cabo la evaluación es la tasa de mala clasificación APER. Como resultado de este proceso se producen dos tasas de error aparente de clasificación incorrecta, una con información proveniente de distribuciones normales sesgadas y otra con la información entendida como testigo en el proceso, para cada combinación de niveles de los factores adoptados.

Figura 1

Efecto de la dimensión, la distancia entre los centroides y el sesgo en una variable, en discriminación lineal parados grupos.

Pasos para la implementación computacional



4. RESULTADOS

Esta sección está dividida en dos partes, con el fin de relatar los hallazgos encontrados como producto de la simulación realizada, de tal manera que la primera presentará los resultados cuando el sesgo aplicado a una de las variables discriminadoras es relativamente pequeño; la segunda sección se dedica a evaluar el efecto en la tasa de clasificación incorrecta cuando el sesgo de la variable es pronunciado. Cada una de las dos secciones se subdivide en dos para contextualizar las situaciones de la regla lineal discriminante: Bajo generación de parámetros (Contexto Poblacional) y complementariamente bajo estimación de los mismos (Contexto Muestral). Para efectos de lectura de los resultados mencionados, las siglas APER_N y APER_SN hacen referencia a las Tasas de Error Aparente de Clasificación Incorrecta calculadas con base en los datos sintéticos generados a partir de las poblaciones normales (N) y normales sesgadas (SN), mientras que las siglas aper_N y aper_SN son las Tasas de Error Aparente de Clasificación Incorrecta en el contexto muestral.

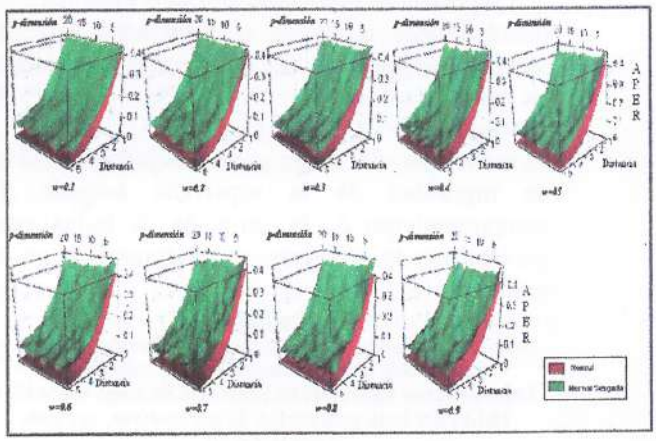
TASAS DE ERROR APARENTE, EN PRESENCIA DE SESGO ENTRE 0.0 Y 0.9 (APER BAJO GENERACIÓN DE PARÁMETROS, CONTEXTO POBLACIONAL)

Cuando se hace variar el sesgo entre 0.0 y 0.9, las APER para las dos grupos de poblaciones se comportan tal como se esperaba, las APER para la discriminación en presencia de sesgo es levemente más grade que las APER para la discriminación sin sesgo, esto para todas las distancias y dimensiones en cuestión. Solo en unos pocos casos sucede lo contrario; estos casos son: Para w = 0.1, 0.2, 0.4, 0.6, 0.7, 0.8, 0.9 cuando p = 11 y D = 4.0. El comportamiento completo para los diferentes valores de sesgo w, se puede apreciar en el siguiente gráfico.

Gráfico 1

Efecto de la dimensión, la distancia entre los centroides y el sesgo en una variable, en discriminación lineal parados grupos.

Tasas de error aparente, en presencia de sesgo entre 0.0 y 0.9 (APER bajo generación de parámetros, contexto poblacional)



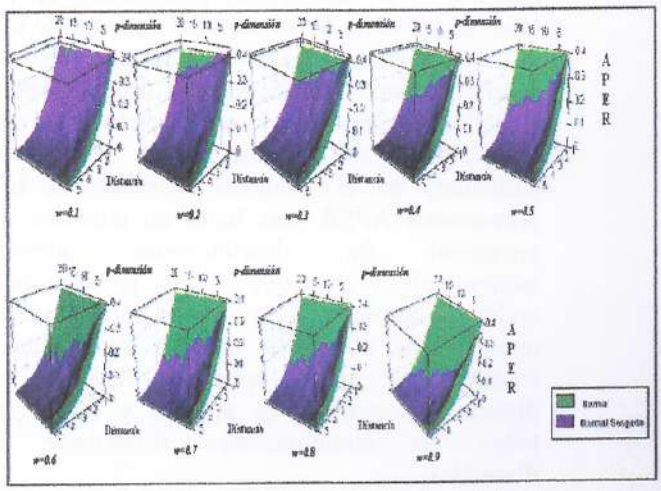
(APER BAJO ESTIMACIÓN DE PARÁMETROS, CONTEXTO POBLACIONAL)

De forma particular, en el contexto muestral, la LDF proporciona APER más bajas en muestras que provienen de distribuciones normales multivariadas sesgadas que en normales multivariadas puras. Este comportamiento se hace más notable a medida que va aumentado la cantidad de sesgo aplicado; iniciando en dimensiones grandes (p grande) y luego para todas las posibles combinaciones de dimensión vs. distancia. Este comportamiento se puede apreciar de mejor manera en la siguiente figura.

Gráfico 2

Efecto de la dimensión, la distancia entre los centroides y el sesgo en una variable, en discriminación lineal parados grupos.

Tasas de error aparente, en presencia de sesgo entre 0.0 y 0.9 (APER bajo estimación de parámetros, contexto muestral)



TASAS DE ERROR APARENTE, EN PRESENCIA DE SESGO ENTRE 1.0 Y 10 (APER BAJO GENERACIÓN DE PARÁMETROS, CONTEXTO POBLACIONAL)

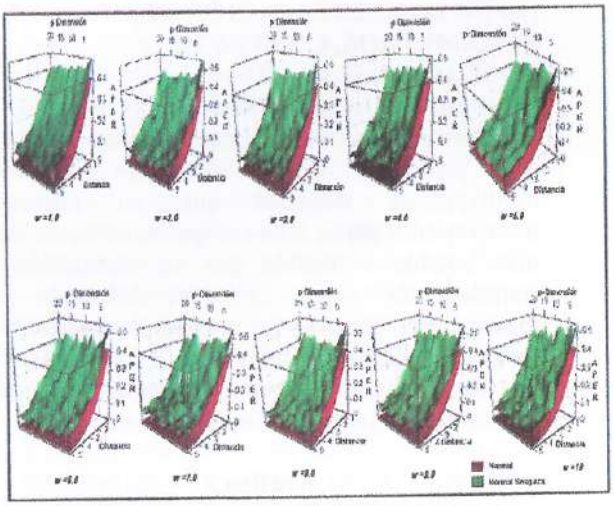
El comportamiento de las APER bajo los factores presentes (sesgo entre 1.0 y 10) no proporcionan evidencia de mejoría, pues aumentar la dimensión no hace que el sesgo baje, por tanto las superficies que representan las APER son más altas para las distribuciones normales multivariadas sesgadas que para las normales puras. Algunas excepciones se presentaron en los niveles (w = 6; p = 9; D = 0.5) y (w = 8, 9; p = 11, D = 0.5).

El incremento del sesgo se ve reflejado en el nivel de rugosidad de la superficie sesgada. El comportamiento de la variación de todos estos parámetros se presenta a continuación en el gráfico 3.

Gráfico 3

Efecto de la dimensión, la distancia entre los centroides y el sesgo en una variable, en discriminación lineal parados grupos.

Tasas de error aparente, en presencia de sesgo entre 1.0 y 10 (APER bajo generación de parámetros, contexto poblacional)



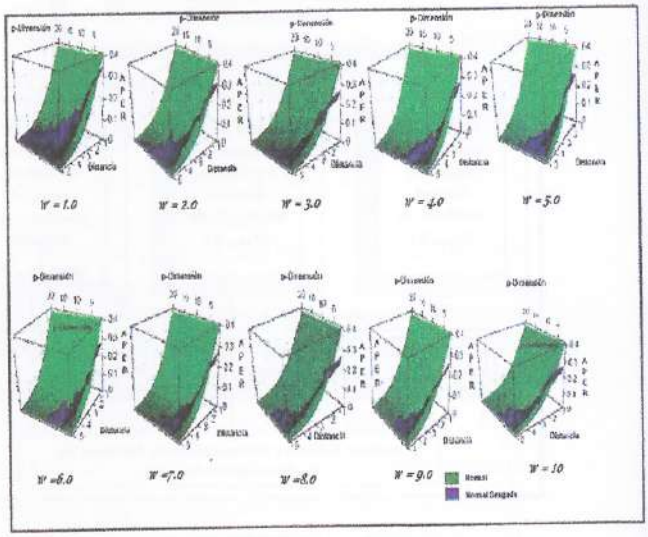
APER BAJO ESTIMACIÓN DE PARÁMETROS, CONTEXTO MUESTRAL)

El comportamiento de las APER del contexto muestral se ve mucho más pronunciado en este caso. El aumento de la cantidad de sesgo w hace que la superficie creada por las tasas de error aparente de mala clasificación se declinen ante la superficie de las tasas de error aparente de mala clasificación sin aplicar sesgo. De forma particular, en el contexto muestral, la LDF proporciona APER más bajas en muestras que provienen de distribuciones normales multivariadas sesgadas que en normales multivariadas puras. Este comportamiento se hace más notable a medida que va aumentado la cantidad de sesgo aplicado; iniciando en dimensiones grandes (p grande) y luego para todas las combinaciones (dimensión vs. distancia).

Gráfico 4

Efecto de la dimensión, la distancia entre los centroides y el sesgo en una variable, en discriminación lineal parados grupos.

Tasas de error aparente, en presencia de sesgo entre 1.0 y 10 (APER bajo estimación de parámetros, contexto muestral)



5. CONCLUSIONES

Las tasas de error aparente de clasificación incorrecta (APER) calculadas en este estudio, en los contextos muestral y poblacional, bajo sesgos (w), desde 0.1 hasta 0.9 y desde 1.0 hasta 10; al variar las distancias Mahalanobis (D), desde 0.5 hasta 5, y variando la dimensión (p), desde 2 hasta 11; fueron consignadas en los gráficos anteriores. De estos resultados a manera de sinopsis se puede sintetizar a partir de la descripción y de lo expuesto en ellos como conclusiones, lo siguiente:

- **Sesgos Menores (w entre 0.1 y 0.9) Contexto Poblacional:**
  - ✓ A pesar de que las diferencias entre las *APER_N* y *APER_NS* son muy pequeñas, la simulación sugiere que éstas aumentan a medida que aumentan la distancia y la dimensión. Para las distancias Mahalanobis D = 0.5, 1.0, 1.5 y 2.0; las superficies APER normales y normales sesgadas son muy cercanas. La forma de las superficies *APER_NS* son en general, más altas y menos homogéneas que las superficies las *APER_N*.
  - ✓ Las *APER_N* son siempre más pequeñas que las *APER_NS*.
  - ✓ Aumentar la distancia y la dimensión entre las poblaciones no atenúa el efecto del sesgo.
- **Sesgos Menores (w entre 0.1 y 0.9) Contexto Muestral:**
  - ✓ Para el sesgo w = 0.1, las diferencias entre las *aper_N* y *aper_N S* son extremadamente pequeñas.
  - ✓ Para valores superiores de w, las *aper_N S* tienden a ser más pequeñas que las *aper_N*. Este comportamiento es más pronunciado en

distancias pequeñas (D = 0.5, 1.0 y 1.5) y dimensiones grandes (p desde 15 hasta 20).

✓ A medida que aumenta w, las *aper_N* son más grandes que las *aper_NS* especialmente a distancias grandes.

- **Sesgos Mayores (w entre 1 y 10) Contexto Poblacional:**

  ✓ Al igual que para sesgos menores, en sesgos mayores se presenta el mismo comportamiento en las APER, pero ésta vez de forma más aguda. Es decir, las *APER_N* son más pequeñas que las *APER_NS* para todas las distancias y todas las dimensiones en cuestión.

  ✓ Las superficies *APER_NS* son mucho más rugosas en este caso. Esto es a causa de la cantidad de sesgo aplicado.

- **Sesgos Mayores (w entre 1 y 10) Contexto Muestral:**

  ✓ Siguiendo el mismo patrón que se presento en sesgos pequeños para el contexto muestral, en sesgos mayores el patrón de tendencia es similar pero más pronunciado.

  ✓ Las *aper_NS* son mucho más pequeñas que las *aper_N* haciendo que la superficie *aper_NS* se esconda debajo de la superficie *aper_N*.

Síntesis:

➢ La simulación permite tener indicios de que el aumento de la dimensión p no mitiga el efecto del sesgo en la tasa de error de clasificación incorrecta, es decir, ésta no se reduce al aumentar el número de variables en el estudio.

➢ La distancia D entre los centroides tiene un efecto manifiesto en las APER pero no atenúa el efecto del sesgo.

➢ La presencia de sesgo debilita la eficiencia de la regla uniformemente para los diferentes valores de D y p.

➢ En el caso particular del comportamiento de la regla plug-in, que merece estudios más puntuales, la presencia de un mayor número de variables con sesgo amerita igualmente estudios específicos.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1]. **RICHARD, A. JOHNSON AND DEAN W. WICHERN.** (2002). *"Applied Multivariate Statistical Analysis"*. Prentice Hall 5 edit.

[2]. **AZZALINI, A. AND VALLE, D.** (1996). *"The Multivariate Skew-Normal Distribution"*. Biometrika, 83, No 4; pp. 715-726.

[3]. **PETER A LACHENBRUCH.** (1997). "*Discriminant Diagnostics*, Biometric", 53, , pp. 1284-1292.

[4]. **MARDIA, K. V; KENT, J. T; BIBBY, J. M.** (1992). "*Multivariate Analysis*. Academic Press", 9 edition.

[5]. **WAHL, PATRICIA W. AND KRONMAL, RICHARD A.** (1977). *"Discriminant Functions when Covariances are Unequal and Sample Size are Moderate"*. Biometrics 33 No. 3, pp. 479-484.

[6]. **ANDERSON, T. W.** (1958). *"An Introduction to Multivariate Statistical Analysis.* New York: John Wiley and Sons, pp 126-137.

[7]. **MCFARLAND AND DONAL P RICHARDS.** (2001). *"Exact Misclassification Probabilities for plug-in Normal Quadratic Discriminant Functions. I The equal means case"*. Journal of Multivariate Analysis, 77, pp. 21-53.