**CAPITULO II**

**2. MARCO TEÓRICO CONCEPTUAL DE LOS MÉTODOS ESTADISTICOS**

**Introducción**

En este capitulo se define la estadística descriptiva y los métodos estadísticos multivariados que se utilizaron en el presente estudio. Dentro de los métodos estadísticos multivariados descritos están: (i) el análisis de componentes principales, el cual detalla, sus características, objetivos y la explicación de la obtención de ellas. El segundo método multivariado definido es (ii) el análisis de conglomerados, en este, se describen los objetivos, las medidas de semejanza, distancia y las técnicas de agrupamiento. Como último método se tiene (iii) el análisis discriminante, en el cual se detalla la obtención de las funciones discriminante para los casos de dos o mas grupos.

**2.1. Estadística Descriptiva**

**Tipos de Curva**

En ocasiones, las frecuencias tienden a acumularse en el lado izquierdo de la grafica con una cola o rama que se extiende hacia la derecha. Se dice que dicha curva es *sesgada* o *asimétrica.* Si la cola o extremidad va hacia la derecha, este tipo de *sesgo* (*o asimetría*) se conoce como *positivo*. Esta condición se denomina como sesgo negativo.

Las curvas pueden clasificarse también con base en su grado de agudización o *curtosis*. Hay tres tipos de curtosis. Cuando la curva es muy aguda y los extremos o cola están mas por encima de la línea de la base. Dicha curva se llama *leptocúrtica.* La curva que es algo achatada se denomina *mesocúrtica,* en tanto que la curva muy aplanada, se llama *platocúrtica*.

**Diagrama de Caja**

Un diagrama de caja es una ilustración gráfica, basada en cuartiles, que ayuda a visualizar un conjunto de datos.

Se requieren cinco tipos de datos para construir un diagrama de caja: el valor mínimo, el primer cuartil, la mediana, el tercer cuartil, y el valor máximo.

**Figura 2.1**

**Diagrama de Caja**



Fuente: N.M. Downie y R.W. Heath, 1986

Autor: Pamela Crow

**Pruebas de Bondad de Ajuste**

Existen diferentes pruebas para verificar el ajuste de los datos a una distribución de probabilidad. Las dos mas utilizadas son el contraste de *X2* de *Pearson* y la *prueba de Kolmogorov-Smirnov*.

**Prueba de Kolmogorov-Smirnov**

Este contraste, que es valido únicamente para variables continuas, compara la función de distribución (probabilidad acumulada) teórica con la observada, calcula un valor de discrepancia, representado habitualmente como *D*, que corresponde a la discrepancia máxima en valor absoluto entre la distribución observada y la distribución teórica, proporcionando asimismo un valor de probabilidad *P*, que corresponde, si estamos verificando un ajuste a la distribución normal, a la probabilidad de obtener una distribución que discrepe tanto como la observada si verdaderamente se hubiera obtenido una muestra aleatoria, de tamaño n, de una distribución normal. Si esa probabilidad es grande no habrá por tanto razones estadísticas para suponer que nuestros datos no proceden de una distribución, mientras que si es pequeña, no será aceptable suponer ese modelo probabilística para los datos.

**2.2. Estadística Multivariada**

## 2.2.1. Introducción

La estadística multivariada es usada para describir y analizar observaciones multidimensionales o multivariadas. Una observación multidimensional se obtiene cuando se releva información sobre varias variables para cada unidad o “individuo” en estudio. La Estadística Multivariada provee herramientas para comprender la relación de dependencia entre variables medidas simultáneamente sobre una misma unidad, para comparar, agrupar y/o clasificar observaciones multivariadas e incluso para comparar, agrupar y clasificar variables. Gran parte de la metodología multivariada se basa en los conceptos de distancia y de dependencia lineal. Las distancias serán usadas como medidas de variabilidad entre pares de puntos que representan los datos multivariados y a partir de ellas es posible analizar similitudes y diferencias entre observaciones y/o variables. Mientras que el análisis univariado explora datos de cada variable independientemente, el análisis multivariado explora tablas de datos de varias variables y por tanto permite contemplar distintos tipos de dependencias entre variables: dependencias entre cada para de variables, entre una variable y todas las restantes, entre pares de variables controlando por el efecto de otras en el sistema multivariado y dependencia conjunta entre todas las variables.

## 2.2.2. Matriz de Datos Multivariados

La organización de datos para un análisis multivariado generalmente se lo realiza en forma de una matriz con n filas, en cada fila se registran observaciones de un mismo individuo, y cada una de las p columnas representa una variable aleatoria. La***Figura 2.2***muestra la matriz de datos multivariados de dimensión nxp. A esta matriz de datos la llamaremos, donde cada fila es un caso u observación multivariada. Una observación multivariada es la colección de mediciones sobre p variables diferentes tomadas sobre el mismo ítem o unidad objeto de estudio.

**Figura 2.2**

**Organización de Datos Multivariados**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| VariableCaso |  |  |  |  |  |  |
| 1 |  |  | ... |  | ... |  |
| 2 |  |  | ... |  | ... |  |
| . | . | . | ... | . | ... | . |
| . | . | . | ... | . | ... | . |
| n |  |  | ... |  | ... |  |

 Fuente: Rencher A., 1998 Autor: Pamela Crow

Cada observación multivariada puede ser representada como un punto en el espacio p por un vector p-dimensional de variables aleatorias con coordenadas igual al valor de cada una de las variables. Este vector de variables puede considerarse como una variable vectorial o multivariante p-variada.

El análisis estadístico multivariado puede ser descriptivo o inferencial; pero esta depende del tipo de variable que se utilice en la investigación. Las variables pueden ser cuantitativas o cualitativas. En las cuantitativas encontramos las variables continuas o intercalares y las discretas; mientras que en las cualitativa, las dicotómicas (binarias) o multinomiales. Observando la matriz de datos multivariados puede observarse que los valores de p variables medidas en n individuos representan la colección de p vectores columnas n-dimensionales (cada columna provee n mediciones de una misma variable).

**2.3. Análisis de Componentes Principales**

El análisis de componentes principales (ACP) es un instrumento de apoyo para otras técnicas multivariadas que permiten la reducción de la cantidad de variables con las que se trabaja.

 Esta técnica multivariada de análisis de datos estudia un número finito de p variables, las cuales constituyen un vector aleatorio en Rp dado por X’ [x1, x2…, Xp], mediante el método estas p variables observadas generan k variables artificiales (siendo k menor que p) las mismas que se pretende, tendrán tanta información como las p variables originales.

El ACP provee una aproximación para la construcción de nuevas variables y para decidir cuántas de estas nuevas variables podrían ser necesarias para representar la información original.

**2.3.1. Características y Objetivo**

Se presenta algebraicamente como una combinación lineal de las p variables aleatorias observadas y geométricamente esta combinación lineal representa la creación de un nuevo sistema de coordenadas obtenidas al rotar el sistema original. Permite describir la estructura de interrelación de variables originales consideradas simultáneamente, determinando así q combinaciones lineales de p variables observables que contengan la mayor parte de la variable total, y así resumir y reducir los datos disponibles.

 El objetivo de la técnica consiste en:

* Reducir el número de variables consideradas tanto como sea posible.
* Encontrar una explicación de los factores que inciden en el comportamiento de las p variables originales.

**2.3.2. Matriz de Varianza - Covarianza**

Las componentes principales, dependen solamente de la matriz de Varianzas-Covarianzas o la matriz de correlación de X1, X2, X3, X4….Xp.

El ACP para ordenar variables se basa en la descomposición espectral de la matriz de covarianza o de correlación de dimensión n×n.

**2.3.3. Autovalores y Autovectores de la Matriz de Varianza - Covarianza**

Los autovalores y autovectores de la matriz de varianzas-covarianzas son usados para obtener las componentes.

Algebraicamente la j-ésima componente principal es una combinación lineal de las p variables originales obtenida como  con j=1,...,p. Las nuevas variables usan información contenida en cada una de las variables originales, algunas variables pueden contribuir más a la combinación lineal que otras. Los coeficientes de cada variable original en la componente son proporcionales al coeficiente de correlación entre la componente y la variable. La varianza de la j-ésima componente principal es  y además se satisface que  para i≠j.

Eligiendo los autovectores como vectores de coeficientes para la combinación lineal se puede demostrar que las componentes principales son combinaciones lineales no correlacionadas cuyas varianzas son tan grandes como sea posible, sujeto a la restricción de que el vector de coeficientes tenga longitud 1. Esta restricción debe ser impuesta ya que de lo contrario la varianza de la combinación lineal podría incrementarse indeterminadamente a través de la multiplicación del vector de coeficientes de la combinación por alguna constante.

**2.3.4. Proporción de la Varianza Poblacional Total**

El numero de componentes principales posibles de construir es p, pero para obtener una dimensión de reducción se selecciona un orden d < p de combinaciones lineales, la cual retendrá una proporción de varianza total no menos del 75% y se usan estas combinaciones como nuevas variables para graficar y analizar los datos sin mayor pérdida de información.

**2.3.5. Obtención de las Componentes Principales**

Para su obtención no se requiere el supuesto de normalidad multivariada, por otra parte, si las componentes principales se derivan de una población normal multivariada se tienen interpretaciones en términos de las elipsoides de confianza.

La técnica de análisis de datos, se basa en el algebra lineal, presenta métodos descriptivos que no hacen ningún tipo de hipótesis probabilísticas, mas bien dan prioridad a la información pero en la búsqueda de interpretación de los factores, se pueden sugerir formulas de hipótesis, a partir de los resultados

**2.4. Análisis de Conglomerado**

El “Análisis de Conglomerados” es una técnica multivariada que se utiliza para agrupar observaciones, variables o entidades de un conjunto de datos en base a sus semejanzas o diferencias. Los objetos pueden corresponder a estructuras identificables como físicas o psicológicas (personas, empresas, países, etc.). Las variables son las características con respecto a las cuales los objetos varían entre sí y que permiten diferenciarlos.

**2.4.1. Objetivo del Análisis de Conglomerado**

El análisis de conglomerado tiene como finalidad ubicar los objetos en grupos o clusters de forma sugerida por los datos, no definidos “a priori”, tal que los objetos en un grupo dado tiendan a ser semejantes en algún aspecto (cohesión interna del grupo) y los objetos en diferentes grupos tiendan a ser distintos (aislamiento externo del grupo). Generalmente es utilizado para conocer el número de grupos y la estructura de estos mismos. Hay otros usos que se le da al análisis de conglomerados como el de clasificación automática, la cual parte de la existencia de un número determinado de grupos y lo que hace es hallar una segmentación razonable de los objetos. También se lo usa para resumir datos o disminuir dimensión más que encontrar grupos “naturales” o “artificiales”, a este procedimiento se lo llama también disección.

**2.4.2. Medidas de semejanza**

En el análisis de conglomerados se parte de una matriz de datos n x p (supongamos p mediciones o variables en cada uno de los n objetos estudiados) que luego es transformada en una matriz de proximidad (n x n) que mide la semejanza o la distancia entre pares de objetos i y j para i,j=1,..., n. Luego se elige un algoritmo de clasificación que define las reglas concernientes al procedimiento de agrupación de los objetos o variables en subgrupos en base a sus proximidades. Algoritmos diferentes se basan en diferentes definiciones de clusters y de semejanzas entre los objetos a agrupar.

2.4.3. Distancias

Dado que el objetivo básico del análisis de conglomerados es medir la asociación entre las entidades a agrupar, es necesario que se establezca una medida de similaridad, o su complemento (medida de disimilaridad). En la formación de grupos, la proximidad está dada por algún tipo de distancia. La selección de una medida de distancia apropiada es fundamental en el uso de cualquier técnica de agrupamiento, sin embargo la selección de esta depende de la naturaleza de las variables ya sean estas: binaria, discreta, continua, de la escala de medición (nominal, ordinal, intervalo, cociente) y del conocimiento del objeto de estudio.

Para los datos con propiedades métricas pueden usarse distancias derivadas de la métrica de Minkoswki, mientras que con datos cualitativos o atributos son más apropiadas medidas de coincidencia o similaridad.

Supongamos que se desea agrupar n observaciones multivariadas, cada una representada por un vector aleatorio p-dimensional. Para medir la distancia entre dos vectores p-dimensionales se utilizan expresiones derivadas de la métrica de Minkowski:

****

Con m=1 se tiene la distancia “Maniatan”. Con m=2 la métrica produce la distancia Euclídea. Incorporando la matriz de varianzas o la matriz de varianzas-covarianzas de las observaciones se derivan a partir de la métrica de Minkowski la distancia estadística o distancia chi-cuadrado y la distancia de Mahalanobis, respectivamente. Sin un conocimiento “a prori” de la estructura de grupos entre las observaciones esas matrices involucrando varianzas y covarianzas son pobremente calculadas. Por esta razón la distancia Euclídea es frecuentemente preferida con motivos de agrupamientos (Johnson y Wichern 1998).

Hay que tener en cuenta que la distancia Euclídea varía con la escala y esta puede ser completamente distorsionada por un simple cambio en ella. Se sugieren que los datos se estandaricen antes de calcular las distancias Euclídeas si es que presenta una varianza muy grande en la escala.

Transformar distancias en medidas de asociación es bastante sencillo, la similitud entre el objeto *i* y *k* es 1 / (1 + dik) si dik es la distancia entre ellos. Sin embargo, lo contrario no es cierto, debido a que las distancias deben satisfacer las condiciones de positividad, simetría y desigualdad triangular. Gower mostró que si la matriz de similitudes es definida no-negativa y la máxima similitud es 1, una medida con propiedades de distancia puede ser obtenida a partir de la similitud como distancias comúnmente usadas entre pares de observaciones con variables.



Cuando se desea agrupar variables más que observaciones, las medidas de similitud más usadas toman la forma de coeficientes de correlación muestral.

2.4.4. Algoritmos de Agrupamiento

En el análisis de conglomerados se utilizan técnicas de clasificación jerárquica y no-jerárquica. En las no-jerárquicas tenemos algoritmos que producen particiones y algoritmos que generan clases no-disjuntas. También existen técnicas como el algoritmo “fuzzy” que produce clases sin superposiciones y otras con cierta probabilidad de superposición.

A partir de n observaciones p–dimensionales se puede construir una matriz n×n de distancias entre las observaciones o una matriz p×p de distancias entre las variables. Como se dijera anteriormente, las matrices de distancia son sometidas a un algoritmo de clasificación para agrupar observaciones y/o variables.

**2.4.4.1. Técnicas de agrupamiento jerárquico**

Las técnicas de agrupamiento jerárquicas están organizados de tal manera que un cluster puede estar contenido completamente dentro de otro cluster, pero no está permitido otro tipo de superposición entre ellos. Los algoritmos de clasificación jerárquicos utilizados con fines de agrupamiento pueden ser acumulativos o aglomerativos y divisorios.

En el caso de los aglomerativos, estos se determinan a través de fusiones de los n objetos/variables por una serie de uniones sucesivas; donde en el inicio hay tantos grupos como objetos y los objetos similares se agrupan primero y esos grupos iniciales son luego unidos de acuerdo a sus similitudes, como las diferencias van disminuyendo, al final todos los subgrupos formarán un solo grupo. Mientras que en el caso de los divisorios particionan los n objetos/variables en subdivisiones cada vez más finas.

Los métodos más utilizados en la práctica de análisis estadístico de datos son los métodos acumulativos o aglomerativos.

Utilizando el procedimiento jerárquico aglomerativo, se muestran el siguiente dendrograma los resultados del agrupamiento, en el que se pueden observar las uniones y/o divisiones que se van realizando en cada nivel del proceso de construcción de conglomerados **(Figura 2.3)**. En el dendrograma se trazó una línea de referencia a nivel de una magnitud de distancia igual a 5, en la cual se pueden identificar 5 conglomerados, si la referencia hubiese estado en 8, se habrían clasificado los objetos en 4 grupos.

**Figura 2.3**

 **Dendrograma Construido por un Procedimiento Jerárquico**

**Aglomerativo de Clasificación**



 Fuente: Rencher A., 1998 Autor: Pamela Crow

Una de las principales características de los procedimientos de agrupamiento jerárquicos aglomerativos es que la ubicación de un objeto en un grupo (cluster) no cambia, o sea, que una vez que un objeto se ubicó en un conglomerado, no se lo reubica, sólo puede ser fusionado con otros objetos pertenecientes a algún otro conglomerado, para formar un tercero que incluye a ambos.

Todos los métodos acumulativos proceden de manera semejante:

1. Cada objeto pertenece a un conglomerado diferente, luego

2. Se fusionan los dos objetos/variables más cercano (conglomerado);

3. Un nuevo objeto/variable se agrega al conglomerado formado por esos dos objetos/variables u otros dos objetos/variables se fusionan formando otro conglomerado.

4. El proceso continúa de manera similar hasta que, eventualmente, se forma un solo conglomerado que contiene todos los objetos/variables como integrantes del mismo.

Las técnicas de agrupamiento jerárquico difieren por las definiciones alternativas de distancia o semejanza que utiliza. Las técnicas acumulativas más comunes son:

**2.4.4.1.1. Encadenamiento Simple (Single linkage)**

Este método utiliza el concepto de mínima distancia y comienza buscando los dos objetos/variables que la minimizan. Ellos constituyen el primer conglomerado. En las etapas siguientes se procede como se ha explicitado en el punto anterior, pero partiendo de n-1 objetos donde uno de ellos es el conglomerado formado anteriormente. La distancia entre conglomerados está definida como la distancia entre sus miembros más cercanos.

Dado que el procedimiento de encadenamiento simple une conglomerados en función de la mínima distancia entre ellos, el procedimiento puede tener problemas cuando hay grupos muy cercanos o con cierta superposición. El procedimiento de encadenamiento simple, es uno de los pocos procedimientos de clasificación que tienen un buen desempeño con configuraciones de conglomerados no-elípticas (datos en cadena)

 **2.4.4.1.2. Encadenamiento Completo (Complete linkage)**

Es un método aplicable tanto para agrupar objetos como variables. La distancia entre conglomerados es la del par de objetos más distantes. Este método es exactamente opuesto al anterior, en el sentido de que las distancias se definen ahora, como la distancia entre pares de individuos más distantes.

**2.4.4.1.3. Encadenamiento Promedio (Average linkage)**

Es un método aplicable tanto para agrupar objetos como variables a partir de distancias o similitudes. Para obtener la distancia entre dos conglomerados, promedia todas las distancias entre pares de objetos donde un miembro del par pertenece a uno de los conglomerados y el otro miembro al segundo conglomerado.

Este algoritmo sigue el mismo procedimiento que los algoritmos de encadenamiento simple y completo excepto que las distancias entre conglomerados se definen como el promedio de distancias entre todos los pares de puntos, con cada miembro del par perteneciendo a uno de los conglomerados del par.

.

**2.4.4.1.4. Otros Procedimientos de Agrupamiento Jerárquicos**

 **Centroide**

Toma el promedio de todos los objetos en un conglomerado (centroide) para representar al conglomerado y medir distancias entre objetos y el conglomerado o entre conglomerados.

**Ward**

Promedia todas las distancias entre los pares de objetos en diferentes grupos, ajustando por las covarianzas.

**2.4.4.2. Procedimientos No-jerárquicos**

Uno de los métodos no jerárquico de agrupamiento más utilizado es el procedimiento conocido como “K-means”. Este procedimiento separa un grupo de objetos en una cantidad elegida de grupos (procedimiento supervisado) haciendo máxima la variación entre conglomerados y minimizando la variación dentro de cada conglomerado.

**2.4.4.2.1. K-means**

El procedimiento comienza con un agrupamiento inicial o con un grupo de puntos semilla (centroides) que formarán los centros de grupo (partición inicial del grupo de objetos en K items). Prosigue asignando cada objeto al grupo que tiene el centroide (media) más cercano. Generalmente se utiliza distancia Euclídea con observaciones estandarizadas o no. Recalcula el centroide para el conglomerado que recibió un nuevo ítem y para el que lo perdió. Repite el accionar descrito varias veces hasta que ya no existen más re-posicionamientos.

**2.5. Análisis Discriminante**

La técnica multivariada de Análisis Discriminante permite describir algebraicamente las relaciones entre dos o más poblaciones de manera tal que las diferencias entre ellas se maximicen o se hagan más evidente.

El Análisis Discriminante requiere el conocimiento de la estructura de grupos de los elementos en estudio, y cada elemento con información para el análisis es clasificado “a priori” del análisis en una de las poblaciones o grupos claramente identificadas.

El análisis discriminante se realiza con dos enfoques diferentes, ya sea con fines predictivos o explicativos.

En el análisis discriminante predictivo esta relacionados a la clasificación, ya sea de nuevas observaciones u observaciones sobre las cuales no se conoce a qué grupo pertenecen. Esta observación nueva (no usada en la construcción de la regla de clasificación), se asignará al grupo en el cual tienen más probabilidad de pertenecer en base a sus características medidas. Para tal asignación es necesario definir reglas de clasificación.

A diferencia del anterior, en el análisis discriminante descriptivo estamos más interesados en las variables empleadas para diferenciar los grupos, y lo que deseamos es determinar cuáles de esas variables son las que más diferencian a los grupos, cuales son importantes y cuales no a efectos de clasificar los sujetos.

**2.5.1. Obtención de las funciones discriminantes**

Técnicamente, se puede decir que el análisis discriminante tratará de encontrar funciones de variables cuyos valores separen o discriminen lo más posible a los grupos existentes. Estas funciones, denominadas funciones o ejes discriminantes, serán combinaciones lineales de las variables originales de la forma:

**Y** = a0 + a1X1 + a2X2  + …. + apXp

donde p es el número de variables explicativas y los coeficientes {a0, a1,…., ap} se eligen de tal forma que se consiga la máxima separación entre los grupos existentes, es decir, tratando de que los valores que toman estas funciones discriminantes Y en los grupos sean lo más diferentes posibles.

Estadísticamente, este criterio equivale a maximizar la varianza “entre grupos” frente a la varianza “dentro de grupos”. Por tanto, los coeficientes {a0, a1,…., ap} se elegirán de tal forma que se consiga maximizar el valor del cociente:

Varianza entre Grupos

Varianza dentro de Grupos

λ

Si la varianza “entre grupos” es grande, es decir si hay grandes diferencias entre los valores que toma la función Y en los distintos grupos, pero la varianza “dentro de grupos” es pequeña, es decir, los valores de Y para municipios de un mismo grupo son muy similares, entonces diremos que la función discriminante separa bien a los grupos, que serán, internamente muy homogéneos y a la vez muy diferentes entre sí.

Hay que señalar que el número de funciones que pueden obtenerse es el mínimo entre el número de variables explicativas disponibles y el número de grupos menos uno. Estas funciones se obtienen de forma sucesiva en función de su capacidad discriminatoria. Así, la primera función discriminante, que será de la forma:

**Y**1 = a01 + a11X1 + a21X2 + …. + ap1Xp

será la que tenga mayor poder discriminatorio, es decir, la que mejor separe los tres grupos. La segunda función, que vendrá definida por:

Y2  = a02 + a12X1 + a22X2 + …. + ap2Xp

Será la siguiente en capacidad discriminatoria y además, estará incorrelacionada con la función anterior Y1.

En ciertas ocasiones, la capacidad discriminatoria de la primera función, Y1, es tan grande, que la información añadida por la segunda función Y2 apenas es relevante y ésta se ignora, ya que su contribución a la separación entre los grupos no es significativa.

El objetivo del análisis discriminante es, precisamente, mejorar la separación entre grupos utilizando no sólo un par de variables, sino todas variables disponibles. Para ello, la información de todas las variables se combina en unas funciones lineales que son las denominadas funciones discriminantes

**2.5.2. Análisis Discriminante de Fisher**

Fisher (1936) presentó la primera aproximación a la clasificación multivariante para el caso de dos grupos. Los coeficientes propuestos por Fisher se utilizan únicamente para la “clasificación”. Al solicitar esta opción se obtiene una función de clasificación para cada grupo. En el caso de dos grupos, la diferencia entre ambas funciones da lugar a un vector de coeficientes proporcional a los coeficientes no tipificados de la función discriminante canónica. Para aplicar estos coeficientes, se calcula cada una de las funciones para un sujeto dado y se clasifica al sujeto en el grupo en el que la función obtiene una puntuación mayor.

Es importante mencionar que el método de Fisher no requiere del supuesto de normalidad multivariada, asume matrices de covarianzas homogéneas entre grupos y usa la métrica de Mahalanobis para la discriminación.

2.5.3. Análisis Discriminante Canónico (ADC)

Cuando más de dos grupos o poblaciones describen la estructura de las observaciones, el método de Fisher es generalizado bajo el nombre de análisis discriminante canónico.

La lógica del ADC para la separación de los grupos se sustenta en la obtención de la combinación lineal (Z) de las variables independientes (Yi), de forma que la correlación entre Z y Yi sea maximizada. La idea básica en el ADC es encontrar los valores de los coeficientes que maximicen la correlación entre Z y Yi. El ADC transforma las variables originales en un número pequeño de variables compuestas, denominadas funciones o variables canónicas, que maximizan la variación entre los grupos y minimizan la variación dentro de ellos.

La combinación lineal para una función discriminante (Z) puede ser representada por:

 Z = µ1 Y1 + µ2 Y2 + ... + µi Yi

Donde µi es el coeficiente canónico y Yi son las variables independientes medidas.

El número máximo de funciones discriminantes canónicas generadas es igual al número de grupos menos uno.

La distancia de Mahalanobis (D2), definida como el cuadrado de la distancia entre las medidas de los valores estandarizados de Z (centros), fue utilizada para verificar si existían diferencias significativas entre los grupos. De esa forma, cuanto mayor el valor de D2, mayor la distancia entre las medias de los dos grupos considerados. El centro de cada grupo representa el valor medio discriminante de los individuos de cada tratamiento. El estadístico lambda de Wilks fue usado para evaluar si las funciones discriminantes canónicas contribuyeron significativamente en la separación de los grupos.