UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERIA
ESCUELA DE INGENIERIA DE SISTEMAS

Informe de Proyecto de Grado:

" Aplicaciones del Análisis Multivariante a dos especies de insectos: Rhodnius Robustus y Rhodnius Prolixus (Insecto transmisor del Mal de Chagas)"

Autor: Doris Maribel Casteletti Avendaño
Tutor: Milagros Van Grieken

Proyecto de grado presentado ante la ilustre Universidad de los Andes como requisito final para optar al titulo de Ingeniero de Sistemas.

MAYO DEL 2000

AGRADECIMIENTO

Quiero expresar mi agradecimiento ante todo a la Universidad de los Andes, casa de estudios, donde me forme como Ingeniero de Sistemas.

A los profesores del Departamento de Investigación de Operaciones por la enseñanza que me brindaron.

Especialmente quiero agradecer a la Profesora Ligia Araque de Tineo quien aportó los datos, tomados del insectario “Pablo Anduze” del Centro de Investigaciones Trujillanas “José Wiltremundo Torrealba” del Núcleo Universitario “Rafael Rangel” de la Universidad de los Andes. Su aporte hizo posible la realización de éste proyecto.

Extiendo mi más sincero agradecimiento al Departamento de Estadística, en especial al Profesor Efrain Entralgo, por su atención y ayuda desinteresada. Sin su guía no hubiera sido posible la consumación de este proyecto.

A Dios, Uno y Trino por ser mi guía y fortaleza.

A María Auxiliadora de los Cristianos, porque ha sido la Madre que me ha protegido, en quien siempre me refugio.

Y a ustedes Papi, Mami y Marisol (manita) porque en el calor del hogar me enseñaron el significado del Amor y la Familia.

RESUMEN

Se expone un resumen de los Métodos del Análisis Multivariante: Análisis de Componentes Principales, Análisis Factorial, Análisis de Clusters, Análisis de Correlación Canónica, Análisis Discriminante. Ofreciendo un soporte al Análisis Multivariante de dos especies de insectos: Rhodnius Robustus y Rhodnius Prolixus involucrados en la transmisión del Mal de Chagas.

Los resultados de este Análisis se obtuvieron a través del paquete estadístico S.P.S.S. (Statistical Program System Software) versión 9.0, el usuario de estas notas debe remitirse a la ayuda que ofrece el S.P.S.S., para familiarizarse con dicho paquete y poder con facilidad repetir la experiencia usando un nuevo conjunto de datos.

DESCRIPTORES: Análisis Multivariado – Investigaciones (Modelo estadístico)
Investigación Operativa.
* QA278
C3

INDICE

CAPITULO III. ESTUDIO COMPARATIVO ENTRE LAS ESPECIES DE INSECTOS: RHODNIUS ROBUSTUS Y RHODNIUS PROLIXUS (INSECTO TRANSMISOR DEL MAL DE CHAGAS)

CAPITULO IV. APLICACIÓN DEL ANALISIS MULTIVARIANTE A DOS ESPECIES DE INSECTOS: RHODNIUS PROLIXUS Y RHODNIUS ROBUSTUS.

INDICE DE FIGURAS

INDICE DE TABLAS

CAPITULO IV. APLICACIÓN DEL ANALISIS MULTIVARIANTE A DOS ESPECIES DE INSECTOS: RHODNIUS PROLIXUS Y RHODNIUS ROBUSTUS.

INDICE DE GRAFICOS

INTRODUCCION

Hoy en día, con el gran desarrollo que ha tenido la microcomputación y con la aparición de paquetes estadísticos sofisticados, elaborados para trabajar en microcomputadoras, el uso de los métodos multivariantes se ha incrementado considerablemente en las diferentes ramas del saber; sin embargo, tanto para el estudiante que se esta formando en el área de estadística, como para el investigador que usa la estadística como una herramienta para su trabajo, la teoría que involucra los métodos estadísticos multivariantes y la interpretación de los resultados, siguen siendo, de cierto modo, una traba para el uso de estas técnicas.

En la actualidad, si bien es cierto que existe en el mercado una buena cantidad de textos excelentes sobre análisis multivariante, la mayoría escritos en ingles; también es cierto, que un porcentaje elevado de estos textos, están dirigidos a estadísticos o personas que tienen buenos conocimientos matemáticos, especialmente de álgebra lineal. Por otro lado, la mayoría de los paquetes estadísticos computacionales existentes, que incluyen temas sobre análisis multivariante, debido a que su objetivo fundamental es facilitar el procesamiento de los datos, solo se dedican a producir una serie de resultados interesantes y no se detienen a explicar su interpretación; por esta razón, muchos investigadores, que a pesar de contar con los paquetes computacionales, siguen teniendo una alta dependencia de los estadísticos, en cuanto a interpretación se requiere.

ESTRUCTURA DE LA TESIS

Capítulo I: Se identifica el proyecto, motivación, objetivos del proyecto y selección de la tecnológica de Desarrollo.

Capítulo II: En él se profundiza en el Análisis Multivariante y sus herramientas.

Capítulo III: Se presenta el estudio comparativo de las especies Rhodnius, una breve reseña histórica, los datos en estudio y una descripción de la enfermedad del Mal de Chagas.

Capítulo IV: Se muestra la aplicación del Análisis Multivariante a dos especies de insectos: Rhodnius prolixus y Rhodnius robustus, mediante el paquete estadístico S.P.S.S. versión 9.0.

CAPITULO I

IDENTIFICACION DEL PROYECTO

I.1. MOTIVACION

En la última década, el tratamiento estadístico de datos multivariados ha sido ampliamente aceptada y aplicada en casi todos los campos de la investigación científica. Muchas razones han justificado el desarrollo de diversas técnicas. Dos de las más importantes son:

1. El análisis estadístico de la data que es necesario realizar en muchas investigaciones científicas.
2. El advenimiento de una computadora de alta velocidad con una amplia capacidad de almacenamiento y el desarrollo de una facilidad disponible y paquetes fáciles de usar de software para implementar los análisis multivariantes de datos.

I.2. OBJETIVO DE LA INVESTIGACION

I.2.1. Objetivo General

Estudiar métodos del Análisis Multivariante y sus aplicaciones.

I.2.2. Objetivos Específicos

- ✓ Indagar y evaluar la factibilidad de aplicación de técnicas de Análisis Multivariante (Análisis: Factorial, de Cluster, Correlación Canónica, Componentes Principales, Discriminante).

- ✓ Facilitar la consulta de los Análisis Multivariante estudiados de manera que puedan ser aplicados a un problema real especifico: Comparación de dos especies de insectos Rhodnius prolixus y Rhodnius robustus.

- ✓ En cuanto a los requerimientos tecnológicos estudiar y manejar la herramienta ha utilizar como paquete estadístico, al S.P.S.S. 9.0 para Windows, para el manejo del Análisis Multivariante.

I.3. SELECCIÓN DE LA TECNOLOGIA DE DESARROLLO

Analizando los requerimientos tecnológicos se ha decidido trabajar, utilizando como paquete estadístico, al S.P.S.S. (Statistical Program System Software) versión 9.0 para Windows.

CAPITULO II

ANALISIS MULTIVARIANTE

En este Capítulo se describe de manera condensada el producto de la investigación bibliográfica sobre Análisis Multivariante y sus herramientas.

II.1. INTRODUCCION AL ANALISIS MULTIVARIANTE

Definición:

El análisis multivariante puede ser definido simplemente como la aplicación de métodos que distribuyen razonablemente un amplio número de mediciones (por ejemplo variables) hechas sobre cada objeto en uno o más campos simultáneamente.

En esta admisión pierde la definición el punto más importante que es el análisis polivariable frente a las relaciones simultáneas entre variables. En otras palabras, las técnicas polivariables difieren de la univariable y bivariable en las que ellos dirigen la atención fuera del análisis del significado y la desviación (variación) de una variable simple, o de las relaciones ambivalentes entre dos variables, para el análisis de tratamientos o correlaciones que reflejan la extensión de relaciones entre tres o más variables. **(DILLON; GOLDSTEIN, 1984)**

Según Kendall ,1980, en el estudio propio del campo multivariado pueden utilizarse diferentes enfoques, tanto por los distintos tipos de situaciones que se presentan al obtener los datos, como por el objetivo específico del análisis. Los más importantes son:

a) **Simplificación de la estructura de los datos.** El objetivo es encontrar una manera simplificada de representar el universo de estudio. Esto puede lograrse mediante la transformación (combinación lineal o no lineal) de un conjunto de variables interdependientes en otro conjunto independiente o en un conjunto de menor dimensión**.**

b) **Clasificación.** Este tipo de análisis permite ubicar las observaciones dentro de grupos o bien concluir que los individuos están dispersos aleatoriamente en el multiespacio. También pueden agruparse variables.

c) **Análisis de la interdependencia.** El objetivo es examinar la interdependencia entre las variables, la cual abarca desde la independencia total hasta la colinealidad cuando una de ellas es combinación lineal de algunas de las otras o, en términos aun más generales, es una función f(x) cualquiera de las otras.

d) **Análisis de Dependencia.** Para ello se seleccionan del conjunto ciertas variables (una o más) y se estudia su dependencia de las restantes, como en el Análisis de Regresión Múltiple o en el Análisis de Correlación Canónica.

e) **Formulación y prueba de hipótesis.** A partir de un conjunto de datos es posible encontrar modelos que permitan formular hipótesis en función de parámetros estimables. La prueba de este nuevo modelo requiere una nueva recopilación de datos a fin de garantizar la necesaria independencia y validez de las conclusiones.

En los casos de poblaciones univariadas, casi siempre es posible caracterizar completamente la distribución de probabilidades a partir de dos parámetros: la media y la varianza. La inferencia estadística exige, entonces,

tomar una muestra aleatoria y calcular los mejores estimadores de estos dos parámetros. El análisis termina con la interpretación de los dos estimadores.

Sin embargo, para el caso multivariado en que se estudia una población p variada, es decir un conjunto de individuos donde se han observado o medido p características o propiedades, se dispondrá de p medias, p varianzas y (1/2)p(p-1) covarianzas, que no solo deben ser estimadas, lo cual no es difícil con las computadoras digitales, sino que *deben ser interpretadas.*

Si se logra una transformación que genere nuevas variables no correlacionadas se eliminan (1/2)p(p-1) parámetros, y si se reducen las dimensiones de p a (p-1), se pasa de (1/2)p(p+3) parámetros poblacionales a ser estimados e interpretados a (1/2)p(p+3) parámetros poblacionales a ser estimados a (1/2)p(p+3) – (1/2)(p-1)(p+2) = (p+1) parámetros a ser estimados e interpretados.

Si bien puede no existir interés en todos los parámetros y, por lo tanto, no es necesario estimarlos, cuanto más sencillo sea el modelo poblacional, más cerca estará el investigador de encontrar una interpretación comprensible de la estructura original mediante la muestra efectivamente observada.

En la **Tabla I** se presenta él número de parámetros estimables en una población multivariada de diferentes dimensiones, el número de parámetros estimables si se efectúa una transformación que genere nuevas variables no correlacionadas. Los datos de la **Tabla I** se representan gráficamente en la **Figura 1**, donde puede apreciarse el crecimiento relativo del número de parámetros estimables.

Tabla I.1. Número de parámetros estimables según el número de variables por considerar.

Numero De Variables P	Numero de Parámetros Estimables		
	Sin Transformar (1/2)p(p+3)	No Correlacionadas 2p	Reduciendo la Dimensión p + 1
1	2	0	0
2	5	4	2
4	14	8	5
6	27	12	7
8	44	16	9
10	65	20	11
20	230	40	21
30	495	60	31

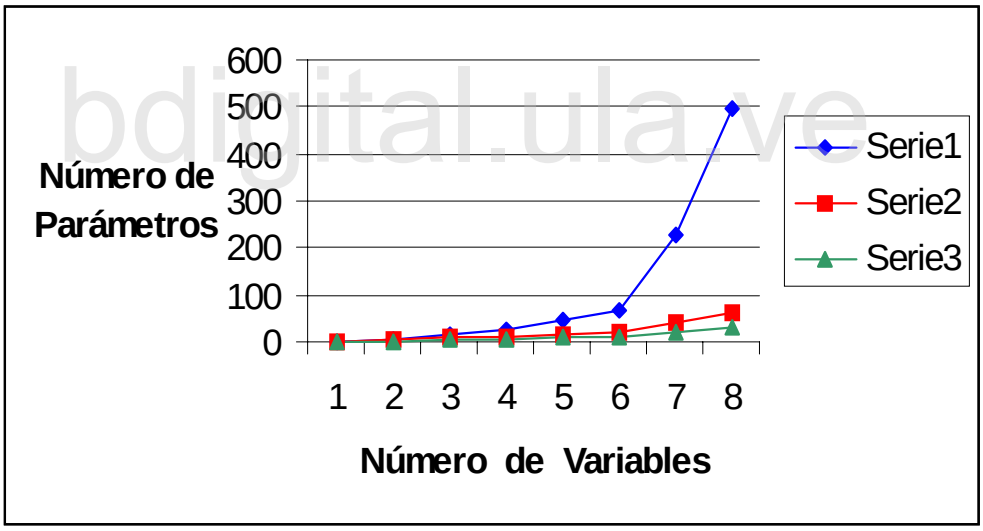


Fig. 1. Número de parámetros estimables en función del número de variables del sistema. El número de parámetros por estimar e interpretar disminuye rápidamente si se efectúa una transformación lineal que genere nuevas variables no correlacionadas; también si se logra disminuir una dimensión.

Cada situación requiere una evaluación particular para utilizar el método de análisis multivariado más adecuado, que permita extraer la máxima información posible del conjunto de datos, pero que a su vez garantice la validez de su aplicabilidad. Las técnicas multivariadas son muy potentes y pueden llevar al investigador a encontrar una justificación (¿por qué no la que “cree” más correcta?) que no se sustente necesariamente en el análisis objetivo de la información recopilada.

Para la mente humana, acostumbrada a pensar y a representar el espacio en dos dimensiones, o a lo sumo en tres, la noción de un multiespacio con cuatro, cinco o p dimensiones resulta difícil de comprender. Hay muchas maneras de acercarse a este concepto y quizás el enfoque matricial y matemático – base del análisis estadístico multivariado- sea él más adecuado.

La medición de varias características de una misma unidad experimental, ya sea en forma simultánea o con ciertos intervalos de tiempo, genera una serie de datos que deben ser analizados con técnicas multivariadas. La unidad experimental puede ser un individuo, una parcela de experimentación, una finca, un animal, una planta, una porción de terreno, y las características serán una serie de atributos, mediciones, evaluaciones, estimaciones, tratamientos o propiedades correspondientes a esas unidades experimentales. No habrá independencia entre las diferentes propiedades utilizadas para caracterizar una unidad y no será posible asignar en forma aleatoria las características, como en un ensayo experimental típico. Habrá, si, independencia entre las unidades experimentales que podrán construir una muestra aleatoria de una población mayor.

Habiendo explicado qué se entiende por universo multivariado, se comprenderá por qué los métodos estadísticos multivariados pueden agruparse en dos conjuntos: los que permiten extraer información acerca de la interdependencia entre las variables que caracterizan a cada uno de los individuos y los que

permiten extraer información acerca de la dependencia entre una (o varias) variable(s) con otra (u otras).

Entre los métodos de análisis multivariado para detectar la interdependencia entre variables y también entre individuos se incluyen el Análisis de Factores, el Análisis por Conglomerados o “Clusters”, el Análisis de Correlación Canónica, el Análisis por Componentes Principales, el Análisis de Ordenamiento Multidimensional (“Scaling”), y algunos Métodos no paramétricos. Los métodos para detectar dependencia comprenden el Análisis de Regresión Multivariado, el Análisis de Contingencia Múltiple y el Análisis Discriminante.

No importa cuan simple o complicado sea un método de análisis de datos; una vez concluida la manipulación algebráica, es necesario interpretar correctamente los resultados obtenidos. En los análisis más difundidos, como el Análisis de Varianza o la Regresión Lineal Simple, luego de lo que para muchos son cálculos tediosos que hoy día pueden hacerse rápida y eficientemente con computadoras de mesa y hasta con calculadoras de bolsillo, hay que interpretar un par de estimadores, por ejemplo: la ordenada en el origen y la pendiente en una ecuación de regresión. Para verificar la hipótesis acerca de los valores encontrados basta generalmente comparar un valor calculado con otro tabulado.

En los últimos cinco años ha aumentado considerablemente él número de textos dedicados al análisis multivariado aplicado a diferentes disciplinas del conocimiento. **(PLA, 1986)**

II.2. HERRAMIENTAS DEL ANALISIS MULTIVARIANTE

Las herramientas que ofrece el Análisis Multavariante para el tratamiento de datos, se describen brevemente a continuación.

Antes se presenta un resumen previo de las herramientas matemáticas y estadísticas básicas, usadas por el Análisis Multivariante.

II.2.1. DATOS Y ANALISIS PRELIMINAR

II.2.1.1. LOS DATOS

La información a procesar con algún método multivariante, generalmente corresponde a uno o varios grupos de datos, donde cada grupo o conjunto consta de varios individuos u objetos y a cada individuo se le ha medido dos o mas variables. Un conjunto de datos que contiene n individuos y p variables, se puede representar como una matriz de orden (n x p), es decir, n filas y p columnas.

En general, un grupo de datos con n individuos u objetos y p variables, sería:

$$X_{n.p} = \begin{array}{c} \text{VARIABLES} \\ \begin{bmatrix} x_1 & x_2 & \dots & x_j & \dots & x_p \\ x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \end{array} \quad \text{INDIVIDUOS} \qquad (1)$$

El elemento x_{ij} denota el valor de la j-ésima variable correspondiente al i-ésimo individuo. La matriz X se puede escribir en cualquiera de las formas siguientes:

$$X = (x_{ij}) = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \left[x(1)x(2)...x(p)\right]$$

i=1,2,...,n

j=1,2,...,n

$$X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ x_{ip} \end{bmatrix} \qquad X_j = \begin{bmatrix} x_{j1} \\ x_{j2} \\ . \\ x_{jn} \end{bmatrix} \qquad (2)$$

Xi denota la i-ésima fila escrita como un vector columna y Xj denota la j-ésima columna.

Con las matrices de datos se pueden hacer comparaciones en dos direcciones: entre filas (individuos) y entre columnas (variables).

II.2.1.2. ANALISIS PRELIMINAR

Antes de emprender la tarea de analizar un determinado conjunto de datos mediante una de las diversas técnicas que nos ofrece el Análisis Multivariante, es muy recomendable formarse una primera impresión u opinión de los datos, haciendo un análisis preliminar que consiste en: calcular y analizar una serie de estadísticas básicas, y realizar algunos gráficos importantes. De ésta manera, se puede observar ciertas características de interés, como lo son los máximos y mínimos, las medidas de tendencia central, las medidas de dispersión, la estructura de las correlaciones entre las variables, posibles relaciones entre variables, posibles tipos de distribución, posible agrupación natural entre observaciones, necesidad de transformación de las variables, detección de valores atípicos, etc.

Otro aspecto que también se puede tomar como parte del Análisis Preliminar, es la prueba de dos supuestos básicos requeridos con mucha frecuencia en los Métodos Multivariantes. Dichos supuestos son:

1. Prueba de normalidad
2. Prueba de igualdad de matrices de covarianzas

II.2.1.3. ESTADISTICAS BASICAS

II.2.1.3.1. Vector de Medias

La media muestra de la j-ésima variable viene dada por:

$$X_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \qquad (3)$$

Si se calcula la media muestral para cada una de las variables, se obtiene el vector de medias muestrales llamado simplemente el Vector de Medias.

$$\overline{\overline{X}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ . \\ \overline{x}_p \end{bmatrix} \qquad (4)$$

En notación matricial, el Vector de Medias se puede expresar como:

$$\overline{\overline{X}} = \frac{1}{n} \sum_{r=1}^{n} Xr = \frac{1}{n} X' 1 \qquad (5)$$

donde:

$$\dot{1} = \begin{bmatrix} 1 \\ 1 \\ . \\ 1 \end{bmatrix}_n$$

II.2.1.3.2. Matriz de Covarianza

El elemento genérico de la Matriz de Covarianza Muestral, S_{ij}, que representa la covarianza muestral entre las variables i-ésima y j-ésima, se define como:

$$S_{ij} = \frac{1}{n-1}\sum_{r=1}^{n}(x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j) = \frac{1}{n-1}\sum_{r=1}^{n} x_{ri}\,x_{rj} - \frac{n}{n-1}\,\bar{x}_i\,\bar{x}_j \qquad (6)$$

$i=1,2,...,n$
$j=1,2,...,n$

Se usa el denominador (n-1) en lugar de n, para hacer que S_{ij} sea un estimador insesgado de la correspondiente covarianza poblacional σ_{ij}. De esta manera, se tiene que la matriz S es un estimador insesgado de $\sum$.

$$S = \begin{bmatrix} S_{11} & S_{12} \ldots S_{1p} \\ S_{21} & S_{22} \ldots S_{2p} \\ . & . \quad . \\ S_{p1} & S_{p2} \ldots S_{pp} \end{bmatrix} \qquad \sum = \begin{bmatrix} \sigma_{11} & \sigma_{12} \ldots \sigma_{1p} \\ \sigma_{21} & \sigma_{22} \ldots \sigma_{2p} \\ . & . \quad . \\ \sigma_{p1} & \sigma_{p2} \ldots \sigma_{pp} \end{bmatrix} \qquad (7)$$

Los términos de la diagonal principal, S_{ii}, corresponden a las varianzas muestrales de las variables.

$$S_{ii} = \frac{1}{n-1}\sum_{r=1}^{n}(x_{ri} - \bar{x}_i)^2 = S_i^2 \qquad i=1,2,...,p \qquad (8)$$

$$S_i = \sqrt{Sii} = \text{Desviación tipica}$$

En notación matricial la Matriz de Covarianzas se expresa de la siguiente forma:

$$S_{ij} = \frac{1}{n-1}\sum_{r=1}^{n}(x_r - \bar{x})(x_r - \bar{x})' = \frac{1}{n-1}\sum_{r=1}^{n} x_r\,x_r' - \frac{n}{n-1}\,\bar{x}\,\bar{x}' =$$

$$= \frac{1}{n-1} X' X = \frac{n}{n-1} \bar{X} \bar{X}' \qquad (9)$$

La covarianza S_{ij}, se puede ver como una medida de asociación lineal entre las variables i,j. Cuando se tiene una covarianza positiva, quiere decir que los valores grandes de la variable i, se corresponden con valores grandes de a variable j y los valores pequeños de i con los valores pequeños de j. Si ocurre lo contrario, es decir, los valores de las variables se mueven en direcciones contrarias, valores grandes de i se corresponden con valores pequeños de j y viceversa, S_{ij} tendrá un valor negativo; si no existe asociación lineal entre las variables, entonces S_{ij} tendrá un valor relativamente pequeño, aproximadamente igual a cero. Debido a que las magnitudes de las covarianzas dependen de las unidades de medida usadas, las covarianzas se hace difíciles de interpretar como medida de asociación lineal entre las variables; para resolver este problema, se definió el coeficiente de correlación muestral que es una versión estandarizada de la covarianza muestral, que no depende de la escala que se use para medir las variables. En otras palabras, el coeficiente de correlación es la covarianza de las observaciones estandarizadas.

II.2.1.3.3. Matriz de Correlación

El término genérico, r_{ij}, de la Matriz De Correlación Muestral, que representa el coeficiente de correlación entre las variables i,j; puede ser visto como una covarianza normalizada para que los valores estén comprendidos entre –1 y 1.

$$r_{ij} = \frac{S_{ij}}{S_i S_j} \qquad \begin{matrix} i=1,2,...,p \\ j=1,2,...,p \end{matrix} \qquad (10)$$

Haciendo todos los cálculos respectivos, se obtiene la Matriz De Correlación Muestral.

$$R = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ . & . & & . \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix} \qquad (11)$$

Si se define,

$$D = \begin{bmatrix} S_1 & 0 & \dots & 0 \\ 0 & S_2 & \dots & 0 \\ . & . & & . \\ 0 & 0 & \dots & S_p \end{bmatrix} \qquad (12)$$

Se puede calcular R y S, mediante las expresiones:

$$R = D^{-1} S D^{-1} \quad \text{y} \quad S = D R D \qquad (13)$$

Un r_{ij} con un valor cercano a 1, indica una asociación lineal fuerte entre variables i y j, con una pendiente positiva; igualmente si el valor esta cerca de –1, la asociación lineal sigue fuerte, pero con una pendiente negativa. Valores cercanos a cero indican ausencia de relación lineal entre las variables, asociación no lineal puede existir y no ser revelada por este estadístico. La no-existencia de correlación rij=0, no implica independencia.

Otro aspecto importante de la Matriz de Correlación, es que su rango ayuda a ver la dimensión efectiva de los datos. Se demuestra (Chatfield, p. 48) que las matrices R y S, simétricas y semidefinidas positivas, tienen el mismo rango que la matriz de datos corregida por la media (X-1X´), que es una matriz de orden n x p. También se demuestra (Chatfield, p. 48) que si $n \leq p$, las matrices R y S son

singulares (rango < p). Si n>p, el rango de R será generalmente igual a p, siempre que no existan restricciones lineales en las variables aleatorias, que hacen que la matriz de datos corregida por la media, sea singular. De esta manera, se dice que el número de restricciones lineales independientes es igual a (p - rango de R).

Por el mismo hecho de tener datos muestrales y debido a errores de redondeo, a pesar de que existen restricciones lineales en las variables aleatorias, los determinantes de R o S generalmente no son iguales a cero, aunque puede ser un valor bastante pequeño. En este caso se dice que la matriz esta mal condicionada, siendo recomendable remover alguna de las variables o reemplazar algunas o todas las variables por un número más pequeño de nuevas combinaciones lineales de las variables. Si la matriz R presenta k autovalores muy pequeños, la dimensionalidad de los datos se puede reducir de p a (p-k).

II.2.1.4. GRAFICOS

Para iniciar esta sección, es oportuno escribir la traducción textual de un párrafo citado en Everitt (p. 1), que dice:

“ El examen preliminar de la mayoría de los datos es facilitado por el uso de diagramas. Los diagramas no prueban nada, pero traen con facilidad a la vista rasgos sobresalientes; ellos no sustituyen las pruebas o test que deben ser aplicados a los datos, pero son muy útiles en sugerir tales pruebas y en explicar conclusiones basadas en ellas.

R. A. Fisher, Statistical Methods for Research Workers”

Los gráficos se han convertido en una ayuda real al investigador, no solo en lo que se refiere al análisis preliminar de los datos, sino en los diferentes métodos multivariantes, donde los gráficos facilitan la tarea de obtener las conclusiones respectivas. Los gráficos en el análisis preliminar ayudaran a buscar relaciones

entre variables, observaciones atípicas, agrupaciones naturales en los datos, a chequear supuestos acerca de distribuciones, a ver la necesidad de transformación de variables, a comparar individuos, etc.

Generalmente, los conjuntos de datos multivariantes tienen más de tres variables, lo cual dificultad obtener un gráfico que englobe todas las variables. Una de las vías de atacar este problema, es graficar las variables en grupos de tres o menos. Sin embargo, existe la limitación de que las distribuciones marginales no necesariamente reflejan la distribución conjunta de todas las variables.

II.2.2. ANALISIS DE COMPONENTES PRINCIPALES

Este es uno de los métodos de análisis más difundidos, que permite la estructuración de un conjunto de datos multivariados obtenidos de una población, cuya distribución de probabilidades no necesita ser conocida.

Se trata de una técnica matemática que no requiere un módelo estadístico para explicar la estructura probabilística de los errores. Sin embargo, si puede suponerse que la población muestreada tiene distribución multinormal, podrá estudiarse la significación estadística y será posible utilizar la muestra efectivamente observada para efectuar pruebas de hipótesis que contribuyan a conocer la estructura de la población original, con un cierto grado de confiabilidad, fijado *a priori* o *a posteriori.*

Los objetivos más importantes de todo ***Análisis por Componentes Principales*** son:

- Generar nuevas variables que pueden expresar la información contenida en el conjunto original de datos.
- Reducir la dimensionalidad del problema que sé esta estudiando, como paso previo para futuros análisis.
- Eliminar, cuando sea posible, algunas de las variables originales si ellas aportan poca información.

Las nuevas variables generadas se denominan *Componentes Principales* y poseen algunas características estadísticas deseables, tales como independencia (cuando se asume multinormalidad) y en todos los casos no-correlación. Esto significa que si las variables originales no están correlacionadas, el Análisis por Componentes Principales no ofrece ventaja alguna.

La literatura acerca de la construcción de los *Componentes Principales*, su uso y sus propiedades es muy amplia. Casi en todos los libros de texto de análisis multivariado se dedica un capitulo al ***Análisis por Componentes Principales***.

II.2.2.1. Origenes

En 1901 Karl Pearson**(24)** publicó un trabajo sobre el ajuste de un sistema de puntos en un multiespacio a una línea o a un plano. Este enfoque fue retomado en 1933 por Hotelling**(18)**, quien fué el primero en formular hasta nuestros días. El trabajo original de Pearson, 1901 **(24)**, se centraba en aquellos componentes, o combinaciones lineales de variables originales, para los cuales la varianza no explicada fuera mínima. Estas combinaciones generan un plano, función de las variables originales, en el cual el ajuste del sistema de puntos es “el mejor”, por ser mínima la suma de las distancias de cada punto al plano de ajuste.

El enfoque de Hotelling se centraba en el ***Análisis de los Componentes*** que sintetizan la mayor variabilidad del sistema de puntos; ello explica quizás el calificativo de “principal”. Por inspección de estos componentes, que resumen la mayor proporción posible de la variabilidad total entre el conjunto de puntos, puede encontrarse un medio para clasificar o detectar relaciones entre los puntos.

Cada punto en el multiespacio p-dimensional es el extremo de un vector X tal que cada uno de sus elementos x(j), para j=1,... ,p, es una medida de la variable j-ésima en un individuo dado. Si se miden n individuos, se obtienen n vectores X y n puntos en el espacio de p dimensiones.

II.2.2.2. Aplicación

Desde sus orígenes, el ***Análisis por Componentes Principales*** ha sido aplicado en situaciones muy variadas: en psicología, medicina, meteorología, geografía, ecología, agronomía.

Dicho análisis se aplica, pues, cuando se dispone de un conjunto de datos multivariados y no se puede postular, sobre la base de conocimientos previos del universo en estudio, una estructura particular de las variables. Cuando se conoce la existencia de una (o varias) variables independientes y, por lo tanto, otro conjunto de variables dependientes, pueden aplicarse las técnicas de regresión múltiple o las de regresión multivariada. Si se sabe que no existe ninguna relación entre las variables (hay independencia o, al menos, no hay correlación), habrá que abstenerse de buscar una explicación de la "relación" entre las variables, o entre los individuos a partir de dichas variables en forma conjunta. En este último caso, en estudios de tipo unidimensional se obtendrán los mismos resultados con técnicas más potentes y en forma menos tediosa, tanto desde el punto de vista computacional como por la facilidad de interpretación.

El ***Análisis de Componentes Principales*** deberá ser aplicado cuando se desee conocer la relación entre los elementos de una población y se sospeche que en dicha relación influye de manera desconocida un conjunto de variables o propiedades de los elementos.

En el ***Análisis por Componentes Principales*** es necesario calcular e interpretar tanto los valores propios generados como los vectores propios. Deberá decidir cuantos valores propios serán considerados si se desea reducir la dimensión original de p variables a m (siendo m<p). Habrá que ser muy cuidadoso al interpretar los vectores propios, ya que el método no es independiente de la escala de medición de las variables originales. **(PLA, 1986)**

II.2.3. ANALISIS FACTORIAL

El ***Análisis Factorial*** es una técnica que consiste en resumir la información contenida en una matriz de datos con V variables. Para ello se identifican un reducido número de factores F, siendo el número de factores menor que el número de variables. Los factores representan a la variables originales, con una pérdida mínima de información.

El modelo matemático del ***Análisis Factorial*** es parecido al de la Regresión Múltiple. Cada variable se expresa como una combinación lineal de factores no directamente observables.

$$X_{ij} = F_{1i}\, a_{i1} + F_{2i}\, a_{i2} + ... + F_{ki}\, a_{ik} + V_i$$

Siendo:

X_{ij} la puntuación del individuo i en la variable j.

F_{ij} son los coeficientes factoriales.

a_{ij} son las puntuaciones factoriales.

V_i es el factor único de cada variable.

Para que el ***Análisis Factorial*** tenga sentido deberían cumplirse dos condiciones básicas: Parsimonia e Interpretabilidad. Según el principio de Parsimonia los fenómenos deben explicarse con el menor número de elementos posibles. Por lo tanto, respecto al ***Análisis Factorial***, el número de factores debe ser lo más reducido posible y estos deben ser susceptibles de interpretación sustantiva. Una buen solución factorial es aquella que es sencilla e interpretable.

II.2.3.1. Antecedentes Históricos

Los antecedentes del ***Análisis Factorial*** se encuentran en las técnicas de regresión lineal, iniciadas por Galton. Un continuador suyo fue K. Pearson (1901),

que presentó la primera propuesta del "*Método de Componentes Principales*", primer paso para el cálculo del ***Análisis Factorial***.

El origen del ***Análisis Factorial*** suele atribuirse a Spearman (1904), en su clásico trabajo sobre inteligencia, donde distingue un factor general (factor G) y cierto número de factores específicos.

Hotelling (1933), desarrolló un método de extracción de factores sobre la técnica de "*Componentes Principales*". Thurstone (1947), expresó la relación entre las correlaciones y las saturaciones de las variables en los factores. Introdujo el concepto de estructura simple. También desarrolló la teoría y método de las rotaciones factoriales para obtener la estructura factorial más sencilla. En un principio las rotaciones eran gráficas. Kaiser (1958) desarrolló el Método Varimax para realizar rotaciones ortogonales mediante procedimientos matemáticos.

A lo largo del desarrollo histórico del ***Análisis Factorial*** se han planteado algunos problemas de fondo que han dado lugar a distintas propuestas de solución. Los aspectos más polémicos han sido:

a- La estimación de las comunalidades.

b- Los métodos de extracción de factores.

c- El número de factores a extraer.

d- Los métodos de rotación de factores.

Por ejemplo, se han propuestos múltiples métodos para la extracción de factores, comprobándose que había distintas soluciones a un mismo problema, según el método que se adoptase. Con esto se plantea el dilema de qué método elegir. Las respuestas han sido distintas según las diversas tendencias. El *Método de Componentes Principales* suele ser el más utilizado. De todas formas hay autores que consideran que el ***Análisis de Componentes Principales*** es distinto ***del Análisis Factorial***.

II.2.3.2. Análisis Factorial vs Componentes Principales

El ***Análisis Factorial*** y el ***Análisis de Componentes Principales*** están muy relacionados.

TABLA II.1. Análisis Factorial vs Componentes Principales

Análisis de Componentes Principales	Análisis Factorial
Trata de hallar componentes (factores) que sucesivamente expliquen la mayor parte de la varianza total. No hace esa distinción entre los dos tipos de varianza, se centra en la varianza total.	Busca factores que expliquen la mayor parte de la varianza común. Se distingue entre varianza común y varianza única. La **varianza común** es la parte de la variación de la variable que es compartida con las otras variables. La **varianza única** es la parte de la variación de las variables que es propia de esa variable.
Busca hallar combinaciones lineales de las variables originales que expliquen la mayor parte de la variación total.	Pretende hallar un nuevo conjunto de variables, menor en número que las variables originales, que exprese lo que es común a esas variables.
No hace tal asunción.	Supone que existe un factor común subyacente a todas las variables.

*En el **Análisis de Componentes Principales**, el primer factor o componente sería aquel que explica una mayor parte de la varianza total, el segundo factor sería aquel que explica la mayor parte de la varianza restante, es decir, de la que no explicaba el primero y así sucesivamente. De este modo sería posible obtener tantos componentes como variables originales aunque esto en la práctica no tiene sentido.*

En resumen tenemos dos grandes tendencias:

a. ***Análisis de Componentes Principales***.
b. ***Análisis Factorial***, dentro del cual existen diferentes métodos.

Ante la variedad de métodos que existen dentro del ***Análisis Factorial***. Kim y Mueller (1978) recomiendan utilizar el de máxima verosimilitud o el de mínimos cuadrados. Sobre la polémica entre Análisis Factorial y Componentes Principales puede consultarse el volumen 25 de Multivariate Behavioral Research (1990).

II.2.3.3. Pasos en el Análisis Factorial

Los pasos que se suelen seguir en el Análisis Factorial son:

1- Calcular la matriz de correlaciones entre todas las variables (conocida habitualmente como matriz R). Examen de esa matriz.
2- Extracción de los factores necesarios para representar los datos.
3- Rotación de los factores con objeto de facilitar su interpretación. Representación gráfica.
4- Calcular las puntuaciones factoriales de cada individuo.

En realidad sólo los dos primeros pasos son indispensables, el 3º y 4º son un complemento.

1- Examen de la Matriz de Correlaciones

El primer paso en el ***Análisis Factorial*** será calcular la Matriz de Correlaciones entre todas las variables que entran en el análisis.

Una vez que se dispone de esta matriz concierne examinarla para comprobar si sus características son adecuadas para realizar un ***Análisis Factorial.*** Uno de los requisitos que deben cumplirse para que el ***Análisis Factorial*** tenga sentido es que las variables estén altamente correlacionadas.

Pueden utilizarse diferentes Métodos para comprobar el grado de asociación entre las variables:

a) El determinante de la matriz de correlaciones

b) Test de Esfericidad de Bartlett

c) Indice de Kaiser-Meyer-Olkin (KMO)

d) Correlaciones Anti-imagen

e) Método de Adecuación de la Muestra (MSA)

f) Correlación Multiple

2- Matriz Factorial

A partir de una Matriz de Correlaciones, el ***Análisis Factorial*** extrae otra Matriz que reproduce la primera de forma más sencilla. Esta nueva matriz se

denomina ***Matriz Factorial***. En la Matriz cada columna es un factor y hay tantas filas como variables originales.

Esos coeficientes pueden interpretarse como índices de correlación entre el factor i y la variable j, aunque estrictamente sólo son correlaciones cuando los factores no están correlacionados entre sí, es decir, son ortogonales. Estos coeficientes reciben el nombre de pesos, cargas, ponderaciones o saturaciones factoriales. Los pesos factoriales indican el peso de cada variable en cada factor. Lo ideal es que cada variable cargue alto en un factor y bajo en los demás.

2.a) Autovalores (Valores Propios)

El cuadrado de una carga factorial indica la proporción de la varianza explicada por un factor en una variable particular.

La suma de los cuadrados de los pesos de cualquier columna de la matriz factorial es lo que denominamos ***Autovalores (Valores Propios)*****,** indica la cantidad total de varianza que explica ese factor para las variables consideradas como grupo.

Las cargas factoriales pueden tener como valor máximo 1, por tanto el valor máximo que puede alcanzar el valor propio es igual al número de variables.

Si dividimos el valor propio entre el número de variables nos indica la proporción (tanto por ciento si multiplicamos por 100) de las varianzas de las variables que explica el factor.

2. b) Número de factores a observar

La Matriz factorial puede presentar un número de factores superior al necesario para explicar la estructura de los datos originales. Generalmente hay un

número reducido de factores, los primeros, que son los que explican la mayor parte de la variabilidad total. Los otros factores suelen contribuir relativamente poco. Uno de los problemas que se plantean, por tanto, consiste en determinar el número de factores que debemos conservar, de manera que se cumpla el ***Principio de Parsimonia.***

Se han dado diversos Criterios para determinar el número de factores a conservar. Uno de los más conocidos y utilizados es el Criterio o Regla de ***Kaise (1960)*** que indicaría lo siguiente: " conservar solamente aquellos factores cuyos valores propios (autovalores) son mayores a la unidad". Este Criterio es el que suelen utilizar los programas estadísticos por defecto. Sin embargo, este Criterio es generalmente inadecuado tendiendo a sobreestimar el número de factores.

Otros Criterios propuestos han sido por ejemplo, *el Scree-test de Cattell (1966)* consistente en representar en un sistema de ejes los valores que toman los Autovalores (ordenadas) y el número de factor (abscisas). Sobre la gráfica resultante se traza una línea recta base a la altura de los últimos Autovalores (los más pequeños) y aquellos que queden por encima indicarán el número de factores a retener.

Vericer (1976) propone el método Minimum Average Partial (MAP) y *Bartlett (1950, 1951)* propone una prueba estadística para contrastar la hipótesis nula de que los restantes p-m.

2.c) Comunalidades

Se denomina *"Comunalidad"* a la proporción de la varianza explicada por los factores comunes en una variable.

La Comunalidad (h) es la suma de los pesos factoriales al cuadrado en cada una de las filas.

El ***Análisis Factorial*** comienza sus cálculos a partir de lo que se conoce como matriz reducida compuesta por los coeficientes de correlación entre las variables y con las Comunalidades en la diagonal.

Como la Comunalidad no se puede saber hasta que se conocen los factores, este resulta ser uno de los problemas del ***Análisis Factorial***.

En el ***Análisis de Componentes Principales*** como no suponemos la existencia de ningún factor común la Comunalidad toma como valor inicial 1. En los otros métodos se utilizan diferentes modos de estimar la Comunalidad inicial:

a) Estimando la Comunalidad por la mayor correlación en la fila i-ésima de la matriz de correlaciones.

b) Estimando la Comunalidad por el cuadrado del coeficiente de correlación múltiple entre x y las demás variables (es el que da el ordenador SPSS por defecto).

c) El promedio de los coeficientes de correlación de una variable con todas las demás.

2. d) Factor Solución

Las técnicas por factor de extracción más usadas y estudiadas son: el ***Método de Factor Principal*** y el ***Método Máximo de Probabilidad*.** El primer método es el más antiguo de los dos y frecuentemente es confundido con el ***Análisis de Componentes Principales*.** EL segundo método es el único Método por extracción de factor que provee corrientemente una base estadística fuerte para probar lo adecuado del modelo de factor analítico básico común.

El ***Método de Factor Principal*** extrae factores de manera que cada uno de ellos cuenta con la cantidad máxima posible de la varianza contenida en un juego de variables siendo factoriadas. Este proceso de extracción es muy parecido al del ***Análisis de Componentes Principales***. La diferencia está en el ***Método de Factor Principal*** donde cada elemento diagonal de la Matriz de Correlación es remplazada por la Comunalidad de variables estimada.

2. e) Interpretación de los factores

En la fase de interpretación juega un papel preponderante la teoría y el conocimiento sustantivo.

A efectos prácticos se sugieren 2 pasos en el proceso de interpretación:

- Estudiar la composición de las cargas factoriales significativas de cada factor.
- Intentar dar nombre a los factores. Nombre que se debe dar de acuerdo con la estructura de sus cargas, es decir, conociendo su contenido.

Dos cuestiones que pueden ayudar a la interpretación son:

- Ordenar la Matriz Rotada de forma que las variables con cargas altas en un factor aparezcan juntas.
- La eliminación de las cargas factoriales bajas (generalmente aquellas que van por debajo de 0.25).

Llamaremos Variable Compleja a aquella que satura altamente en más de un factor y que no debe ser utilizada para dar nombre a los factores.

Factores Bipolares, son aquellos factores en los que unas variables cargan positivamente y otras tienen carga negativa.

3- Rotaciones Factoriales

A partir de la Matriz Factorial muchas veces resulta difícil la interpretación de los factores. Para facilitar la interpretación se realizan lo que se denominan ***Rotaciones Factoriales***.

La ***Rotación Factorial*** pretende seleccionar la solución más sencilla e interpretable. En síntesis consiste en hacer girar los ejes de coordenadas, que representan a los factores, hasta conseguir que se aproxime al máximo a las variables en que están saturados.

La saturación de factores transforma la Matriz Factorial inicial en otra denominada ***Matriz Factorial Rotada***, de más fácil interpretación. La ***Matriz Factorial Rotada*** es una combinación lineal de la primera y explica la misma cantidad de varianza inicial.

Como hemos dicho el objetivo de la rotación es obtener una solución más interpretable, una forma de conseguirlo es intentando aproximarla al ***Principio de Estructura Simple (Thurstone, 1935).*** Según este principio, la Matriz Factorial debe reunir las siguientes características:

1- Cada factor debe tener unos pocos pesos altos y los otros próximos a 0.

2- Cada variable no debe estar saturada más que un factor.

3- No deben existir factores con la misma distribución, es decir, los factores distintos deben presentar distribuciones de cargas altas y bajas distintas.

Estos tres principios no suelen lograrse, lo que se trata es de alcanzar una solución lo más aproximada posible a ello.

Con la ***Rotación Factorial*** aunque cambie la ***Matriz Factorial*** las Comunalidades no se alteran, sin embargo, cambia la varianza explicada por cada factor.

Existen varios ***Métodos de Rotación*** que podemos agrupar en dos grandes tipos: ***Ortogonales y Oblicuos***. La más recomendable es la ***Rotación Ortogonal***, aunque en el caso de que existan razones para pensar que los factores están correlacionados entonces utilizaremos la ***Rotación Oblicua***.

De entre las ***Rotaciones Ortogonales*** la más utilizada es la ***Varimax*** mientras que en las ***Oblicuas*** **es la** ***Oblimin*****. (1, INTERNET)**

II.2.4. ANALISIS DE CLUSTERS O ANALISIS POR CONGLOMERADO

En este punto estamos principalmente interesados en descubrir las relaciones mutuas que existen entre las variables. Los objetos sobre los cuales fueron tomadas las medidas fueron asumidos como homogéneos. Para la mayor parte no había interés, en la posibilidad de que un juego dado de objetos podría agruparse en tópicos que desplegaron diferencias sistemáticas. Sin embargo, muchos marcos de la aplicación, hay razón para creer que el juego de objetos puede clasificarse en subgrupos que difiere de maneras significativas. El término usado comúnmente para la clase de procedimientos que buscan separar los datos del componente en los grupos es el ***Análisis de Cluster***.

La búsqueda para clasificaciones o tipologías de objetos o personas es natural a una amplia variedad de disciplinas. Aunque inicialmente primitivo, el campo creció al confiar en las técnicas numéricas más objetivas que se hicieron posible por el advenimiento de la computadora de gran velocidad que tiene capacidades del almacenamiento enormes. Así, hoy nosotros vemos aplicación de ***Análisis de Cluster*** a diversas áreas tales como:

(1) La psicología - clasificando a los individuos en los tipos de personalidad;

(2) El análisis regional - clasificando ciudades en tipologías basadas en variables demográficas y fiscales;

(3) Comercializando investigación - clasificando a clientes en segmentos en base a los factores psicográficos y uso del producto;

(4) La química - la clasificación de compuestos basados en sus propiedades de actuación; y así sucesivamente.

II.2.4.1. Apreciación global de Análisis de Cluster

La **Figura 2** proporciona una apreciación global del procedimiento de ***Análisis de Cluster***. El proceso empieza típicamente tomando, dice, medidas de p en cada uno de los n objetos. La matriz n x p de datos crudos se transforma entonces en una matriz n x n de similitud o, alternativamente, medidas de distancia, donde se computan las similitudes o distancias entre los pares de objetos a través de las variables de p. Luego un algoritmo de Cluster es seleccionado, definiendo las reglas donde Cluster involucran los objetos en subgrupos en base a similitudes encierre-objeto. Como indicamos, la meta en muchas aplicaciones de Cluster es llegar a los objetos de este que muestran una pequeña variación relativa dentro-Cluster y la variación entre-Cluster. Como un paso final, los Cluster descubiertos están contrastados en los términos de sus valores principales en las variables de p u otras características de interés.

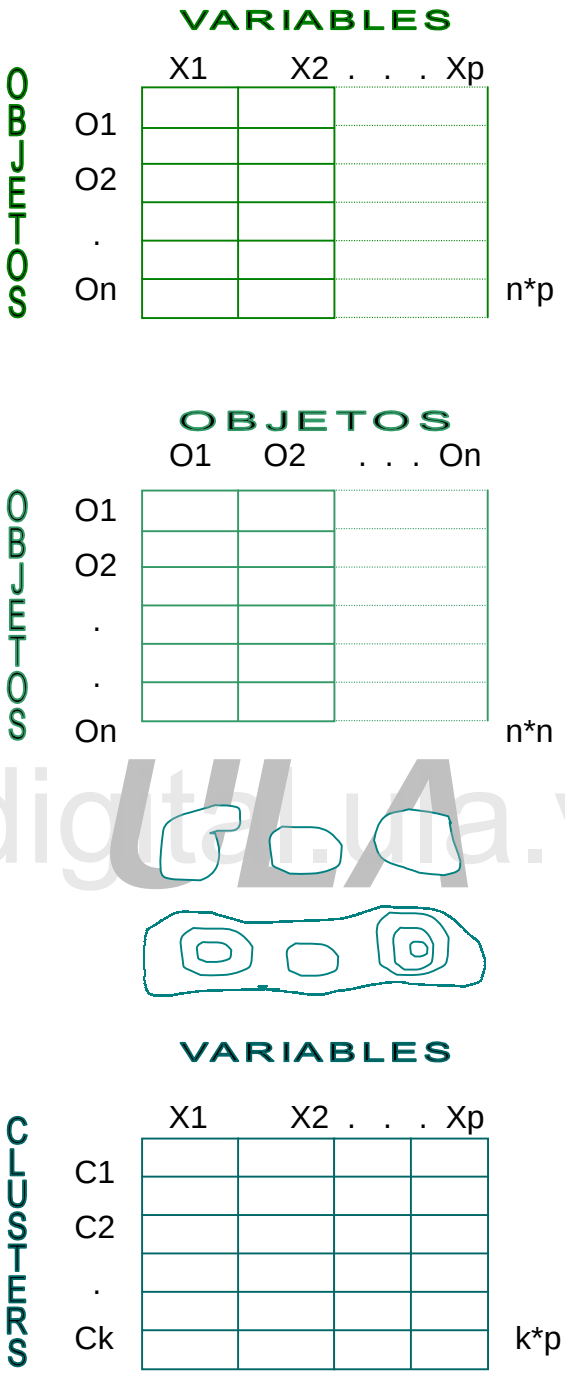


Fig. 2. Apreciación Global del Procedimiento del Análisis de Cluster.

Hay dos problemas importantes en la aplicación del procedimiento de Cluster descrito anteriormente. Primero, necesitamos decidir en una medida de similitud del entierro-objeto; pero esto requiere que definamos lo que es el significado por similitud que no siempre es fácil. Segundo, debemos especificar un procedimiento para formar los Cluster, basado en la medida de similitud escogida. Sin embargo, debemos señalar, que hay docenas de medidas de similitud que se han usado literalmente en las aplicaciones de Cluster, y una lista aparentemente interminable de posibles algoritmos de Cluster. Naturalmente, nuestra discusión será en ciertos casos incompleta. Discutiendo varias de las medidas de similitud más populares y el criterio de Cluster esperamos proporcionar un tratamiento representativo e informativo.

No hay ningún acuerdo universal en lo que realmente constituye un Cluster. Principalmente, el significado de tales términos como Cluster o similitud dependen finalmente de los juicios de valor del usuario. En nuestra discusión de ***Análisis de Cluster*** adoptaremos una definición intuitiva de Cluster que enfatiza en cómo podrían descubrirse visualmente Cluster en dos o tres dimensiones.

Para ilustrar, consideramos los objetos como en un espacio p dimensional, con cada uno de las variables de p representadas por uno de los ejes del espacio. Un sistema de coordenada p dimensional se define ahora en el espacio por los valores de las variables para cada objeto. Nosotros podemos describir Cluster como regiones continuas que aparecen en el espacio como una masa relativamente grande, es decir, una alta densidad de puntos que están separados de otras regiones por regiones que tienen una masa relativamente pequeña (una baja densidad de puntos). El término Cluster natural es usado frecuentemente para describir Cluster basados en este tipo de razonamiento.

Los Cluster naturales no imponen restricciones a priori en la estructura de los datos. Por consiguiente, los datos permiten dictar los modelos de Cluster

encontrados. Esto es opuesto a otras definiciones propuestas. Por ejemplo, definiciones que confían en lo que sólo podría llamarse criterio de encierro - es decir, los objetos en un Cluster deberían encerrarse en cada uno como los objetos en otros Cluster - puede ser restrictivo en el sentido que pueden tener dificultad identificando otros Cluster como aquéllos de la variedad esférica. Las **Figuras 3** y **4** muestran dos juegos de datos que exhiben estructuras diferentes. Los presentes Cluster estan, pensamos, bastante obvios. En general, la mayoría de las técnicas de Cluster no tendrían problema descubriendo la esférica-tipica de Cluster mostrada en la **Figura 3**. Sin embargo, procedimientos que dan énfasis al criterio de encierro tendrían un tiempo difícil recuperando los Cluster naturales mostrados en la **Figura 4**.

Fig. 3. Esférica-tipica de Cluster **Fig. 4. Cluster naturales**

II.2.4.2. Criterio de Cluster

Como indicamos, el investigador que quiere realizar un ***Análisis de Cluster*** se enfrenta con lo que parece ser una lista interminable de los algoritmos de Cluster para escoger entre ellos. La mayor parte, los algoritmos de Cluster dependen en tecnología de la computadora de gran velocidad para la eficacia de los cómputos y se esfuerzan por encontrar algún criterio que esencialmente aumente al máximo la variación de entre-Cluster a la variación del dentro de-Cluster.

Como muestra la **Figura 5,** la variación del entre-Cluster puede ser juzgada evaluando la distancia entre los centros del Cluster en comparación con la distancia de un miembro del Cluster a un centro del Cluster (Note que para la facilidad de presentación se muestran los Cluster en la figura en dos dimensiones, mientras realmente los Cluster existen un espacio multidimensional). Para entender cómo puede usarse este criterio para formar Cluster, imaginemos empezando con un número dado de centros del Cluster escogido arbitrariamente o en juicio, y asignando objetos al centro del Cluster más cercano, computando el medio o centro de gravedad del Cluster resultante, y haciendo malabares de un lado a otro, entonces los objetos entre los Cluster, recalculando cada tiempo de los centros de gravedad y el resultante entre-Cluster y la variación del dentro de-Cluster hasta la proporción es suficientemente grande.

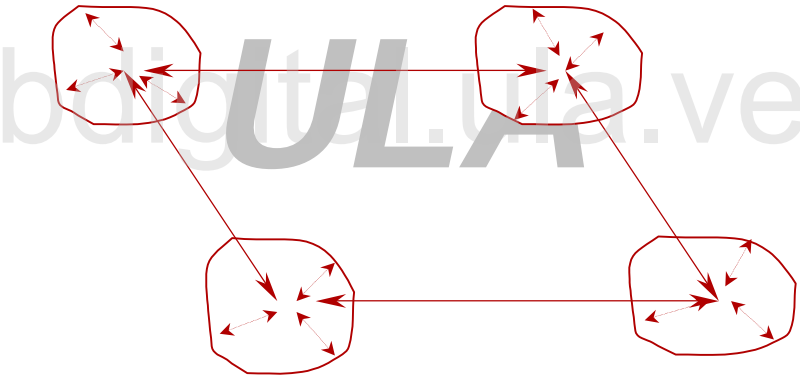


Fig. 5. La variación del entre-Cluster puede ser juzgada evaluando la distancia entre los centros del Cluster en comparación con la distancia de un miembro del Cluster a un centro del Cluster.

II.2.4.3. Relación con otras técnicas

Tabla II.2. Análisis de Clusters vs. Análisis de Componentes Principales.

Análisis del Cluster	Análisis de Componentes Principales
Identifica un número más pequeño de grupos tales como elementos que pertenecen a un grupo dado que son, en algún sentido, más similares a cada uno de los elementos que pertenecen a otros grupos.	El objetivo es reducir el juego original de variables puestas en correlación a algún juego ortogonal más pequeño de compuestos lineales
Intenta reducir la información de un juego completo de n objetos a una información general, digamos, subgrupos de g, donde g <n.	Es una técnica de Cluster donde el enfoque está en las columnas de la matriz de los datos, eso es, las variables.
Puede ser considerado como otra técnica para la reducción de datos.	
Vemos el análisis de Clusters como un enfoque en filas, esto es, objetos o individuos de la matriz de datos donde el objetivo es reducir el número de los distintos tipos de entidades agrupadas en Clusters.	

Aunque al parecer muy similar, es importante entender las diferencias entre el ***Análisis del Cluster*** y ***Análisis Discriminante*****:**

Tabla II.3. Análisis de Clusters vs. Análisis Discriminante.

Análisis de Cluster	Análisis Discriminante
En el ***Análisis del Cluster*** se empieza con grupos que no son inicialmente diferenciados y se pregunta si un grupo dado puede dividirse en subgrupos que difieren de manera significativa.	En el ***Análisis Discriminante*** hay una asunción tácita donde los grupos son conocidos a priori; es decir, se asume que todas las observaciones son clasificados correctamente a la salida. En muchos tópicos de la investigación, sin embargo, el científico social está inseguro de las agrupaciones naturales que podrían estar presentes. En tales casos, el investigador no puede tener ninguna opción sino confiar en las medidas disponibles con el propósito de decidir si las observaciones entran en grupos del elector, y en ese caso, para delinear los grupos. Así, en el ***Análisis Discriminante*** nosotros empezamos con grupos a priori bien-definidos preguntando cómo los grupos dados difieren.

Una vez que los objetos han estado en Cluster, hay una necesidad de comparar los varios Clusters para conseguir una idea de cómo difieren. Un acercamiento sincero, simple es comparar los Clusters con respecto a sus medias y variaciones en el p computando las similitudes del entierro-objeto o en el juego de variables externas para las que la información está disponible en los miembros de Cluster pero que no se usó en tal procedimiento de Cluster.	También pueden derivarse los Clusters de un ***Análisis Discriminante***. El ***Análisis Discriminante*** podría determinar qué variables estaban contribuyendo mayormente a las diferencias del perfil significante entre los Clusters y, además, que mantenga un vehículo prediciendo el número de miembros del Cluster de una muestra futura de objetos.

(**2, Dillon, William R.; Goldstein, Matthew)**

II.2.5. ANALISIS DE CORRELACION CANONICA

En muchos tópicos de la investigación, el científico social encuentra un fenómeno que no es mejor descrito en términos de un solo criterio sino, debido a su complejidad, en términos de un número medido de respuestas. En tales casos, el interés puede centrarse en la relación entre el juego de medidas del criterio y el juego de factores explicativos. En investigaciones psicológicas, por ejemplo, podríamos estar interesados en la relación entre un juego de variables de personalidad, por un lado, y varias medidas de habilidad por el otro lado. En los campos comerciales o económicos, podríamos estar interesados en la relación entre un juego de índices de precio y un juego de índices de producción, prediciendo uno y otro. El estudio de la relación entre un juego de variables del predictor y un juego de medidas de la contestación es conocido como ***Análisis de la Correlación Canónico***.

II.2.5.1. ¿Qué es el Análisis de la Correlación Canónico?

En general, el ***Análisis de la Correlación Canónico*** describe una técnica estadística multivariada que investiga la relación entre dos juegos de variables. En la mayoría de las aplicaciones, sin embargo, los dos juegos de variables no se tratan simétricamente; más bien, un juego es el predictor, es decir, el juego de variables independientes, y el otro es el juego de medidas del criterio. Revoque eso en *Regresión Múltiple* que nuestro acercamiento era para encontrar una combinación lineal de las variables del predictor originales que mejor explicaron la variación en la medida del criterio. En el ***Análisis Canónico*** la idea es casi la misma, sólo que ahora buscamos dos combinaciones lineales, uno para el predictor y uno para el criterio, tal que su correlación del producto-momento ordinaria es tan grande como posible.

En un ***Análisis de variantes Canónico*** están computados ambos juegos de variables. Una variante es análoga a una dimensión o factor en un ***Análisis de los Componentes Principales***. La diferencia es que una variante consiste en un

predictor máximamente puesto en correlación y una parte del criterio. Un máximo de variantes de M puede extraerse, donde M es el número de variables en el juego más pequeño. Como en el ***Análisis de los Componentes Principales***, las variantes de M se extraen para que sean independientes unos de otros.

II.2.5.2. Cuándo Usar Análisis de la Correlación Canónico

El ***Análisis de la Correlación Canónico*** podría usarse analizando varias variables del predictor y variables del criterio simultáneamente. Es particularmente apropiado cuando las variables del criterio están relacionadas entre ellas. En tales casos puede descubrir relaciones complejas que reflejan la estructura entre las variables predictoras y de criterio. Esto aparece en el ejemplo anterior donde el ***Análisis de la Correlación Canónico*** se presenta como una técnica estructural y funcional que el juego predictor y de criterio, siendo estructurada para producir una máxima correlación entre los juegos.

Sólo cuando una variable del criterio está disponible, el ***Análisis Canónico de la Correlación***, reduce el ***Análisis de Regresión Multiple.*** Esto plantea la pregunta: ¿por qué el Análisis de la Regresión Múltiple no se realizan separadamente, uno para cada variable del criterio? Esta aproximación no es recomendada. El ***Análisis de Regresión*** separado frustra el propósito de tener un criterio de medidas múltiple, desde que la información suministrada por la correlación entre las variables del criterio que no es tomada en cuenta.

II.2.5.3. Los Requisitos de los datos y suposiciones para el Análisis Canónico

Usar el ***Análisis de Correlación Canónico cuidadosamente*** para los propósitos descriptivos no requiere ninguna suposición distribucional. En tales casos, los predictores y variables del criterio pueden medirse a nivel nominal u ordinal. Para probar la importancia de las relaciones entre las variantes canónicas, sin embargo, los datos deben reunir los requisitos de normalidad del multivariate y homogeneidad de variación. (**2, Dillon, William R.; Goldstein, Matthew)**

II.2.6. ANALISIS DISCRIMINANTE

El término “discriminación” fue introducido por R.A. Fisher, en 1936, en el primer tratamiento de problemas de separación de grupos.

Con el Análisis Discriminante se analiza la información cuantitativa comparándola con una variable “grupo” definida a priori. En el Análisis Discriminante se determina si el grupo de variables observadas es adecuado, o no, en la separación de los grupos definidos a priori.

Dentro del contexto del análisis de datos, en la mayoría de las áreas del conocimiento, existen líneas de investigación donde el análisis discriminante tiene sus aplicaciones; por ejemplo:

En un centro de rehabilitación para alcohólicos y drogadictos, donde los pacientes llegan voluntariamente en busca de solucionar su problema, existe un tratamiento para alcohólicos basado primeramente en un proceso de desintoxicación y luego en sesiones de terapia tanto individual como de grupo; el paciente tiene que vivir en el centro durante dos meses que dura el tratamiento. Antes del comienzo y después de terminar el tratamiento, el paciente es sometido a un test psicológico de inventario de personalidad, donde se le miden una serie de variables. En base a los resultados obtenidos en estas pruebas, con el objetivo de diseñar un programa de control al paciente después que deja el centro, al especialista le gustaría clasificar cada paciente en uno de los tres grupos siguientes:

Grupo 1: No vuelven a tomar después del tratamiento.

Grupo 2: Toman esporádicamente.

Grupo 3: Siguen tomando como antes.

Un parasitologo podría estar interesado en el estudio de las diferencias que existen entre las distintas especies del transmisor de la leishmaniasis, basándose

en las diferentes características del transmisor, tales como medidas morfológicas, comportamientos, hábitat, etc. De esta manera, el investigador podría diferenciar entre especies muy afines o parecidas, o incluso, hasta identificar nuevas especies.

Un conjunto de enfermedades pueden tener una serie de síntomas muy parecidos, por lo que al especialista le sería de gran ayuda contar con un mecanismo fácil y sin traumas para el paciente, que le permitiera, basándose en los síntomas que presenta, dar un diagnostico tentativo, con cierto grado de precisión, de la enfermedad que padece un determinado paciente.

Sobre la base de un conjunto de variables socioeconómicas, al Gerente de Crédito de una entidad financiera, como ayuda para la toma de decisiones en el otorgamiento de créditos, le gustaría clasificar los créditos de los deudores potenciales, en créditos de alto riesgo, de bajo riesgo y sin riesgo aparente.

Supóngase que para cada uno de los ejemplos anteriores, el investigador cuenta con información previa o experiencia acumulada; por ejemplo, el centro de rehabilitación ha registrado el seguimiento que se le ha hecho a cada paciente después de terminar con el tratamiento, es decir, el centro observa al paciente durante los seis meses siguientes al tratamiento y lo clasifica en uno de los tres grupos, de acuerdo al comportamiento; el parasitologo posee información de un numero suficientemente grandes de transmisores, que ya han sido clasificados correctamente dentro de las distintas especies; el especialista, para un numero considerable de pacientes, tiene un registro apropiado de todos los síntomas de pacientes que fueron correctamente diagnosticados; el gerente de crédito cuenta con el registro de ciertas variables claves, de todos los clientes que les han otorgado créditos. Si este es el caso, el análisis discriminante puede ser de gran ayuda en el logro de lo que se plantea en cada uno de los ejemplos.

El análisis discriminante comprende toda una metodología estadística multivariante, orientada fundamentalmente a alcanzar dos objetivos básicos:

1. Explorar las posibles diferencias que existen entre dos o más poblaciones y analizar la naturaleza de dichas diferencias. En este sentido, se trata de hallar funciones discriminantes, que dependan de las variables originales y que separen las poblaciones tanto como sea posible.
2. Conociendo ciertas características básicas de grupos de individuos pertenecientes a dos o más poblaciones determinadas, construir criterios o reglas de clasificación, que permitan asignar un objeto o individuo "desconocido" a una de las poblaciones conocidas.

En la práctica, una función que separa adecuadamente poblaciones puede servir como base para definir una buena regla o criterio de clasificación, y viceversa, una buena regla de clasificación puede dar origen a un buen procedimiento de separación.

Supóngase que $f_1(X)$, $f_2(X)$,...,$f_g(X)$ son funciones de densidad definidas en R^p y correspondientes a las poblaciones π_1, π_2,...,π_g, $g \geq 2$, de tal manera que un individuo u objeto perteneciente a la j-ésima población multivariante (π_j), y cuyo vector de medidas es denotado por X, tenga funciones de densidad $f_j(X)$. Debido a que uno de los objetivos del análisis discriminante, es clasificar un individuo "desconocido" como perteneciente a una de las poblaciones, basándose en el vector de medidas; la regla discriminante o criterio de clasificación d, se basará en una participación de R^p en las regiones R_1,R_2,...,R_g mutuamente excluyentes y exhaustivas ($\bigcup_{i=1}^{g} R_i = R^p$), de forma tal que el individuo X se asignará a la población π_j si $X \in R_j$, j=1,2,...,g. En la medida que $\int_{Rj} f_j(X)d_x$ tienda a uno, j=1,2,...,g, más precisa será la discriminación. Puede ser que para alguna población exista una

mayor probabilidad de ocurrencia debido a que el tamaño de ésta sea relativamente mayor que las demás; cuando esto ocurre, se pueden tomar en cuenta estas probabilidades de ocurrencia o probabilidades a priori, para la construcción de las reglas de clasificación (enfoque Bayesiano), las probabilidades a priori se denotan por $p_1,p_2,...,p_g$, tal que $\sum_{i=1}^{g} p_i=1$.

En análisis discriminante se pueden considerar un conjunto de situaciones, que dependen básicamente de supuestos que se hacen acerca de las distribuciones poblacionales; así por ejemplo, se tiene:

1. Una situación poco común en la práctica, pero la más oportuna para los desarrollos teóricos, es cuando las funciones de densidad $f_j(X)$, j=1,2,...g, son completamente conocidas, es decir, se conoce tanto la forma de la distribución como los parámetros involucrados.
2. Muchas veces se conoce la forma de la distribución pero se desconocen los valores reales de los parámetros; en estos casos lo que se hace es estimar dichos parámetros, partiendo de datos muéstrales.
3. También existe un enfoque empírico, donde no se asume forma particular alguna para las distribuciones de las poblaciones.

Una vez construida una regla de clasificación, independientemente del método usado, esta regla no garantiza una clasificación libre de errores; por ejemplo, si se esta clasificando entre dos poblaciones, es posible que un determinado sujeto sea clasificado en la población π_1 cuando en realidad pertenece a la población π_2, o viceversa. Una de las vías de medir la “bondad” de un criterio o regla de clasificación, es mediante las probabilidades de clasificación errónea (clasificación equivocada) de ese criterio, por ejemplo, la probabilidad condicional P(1/2), de clasificar un individuo como proveniente de la población π_2 cuando en realidad pertenece a π_1, es:

$$P(1/2) = P(X \in R_2/\pi_1) = \int_{R_2} f_1(X)dx \qquad (1)$$

Similarmente, la probabilidad condicional P(1/2), de clasificar un individuo como de π_1 siendo en realidad de la población π_2, es:

$$P(1/2) = P(X \in R_1/\pi_2) = \int_{R_1} f_2(X)dx \qquad (2)$$

De la misma manera se define la probabilidad de clasificar correcta dentro de una determinada población.

$$P(i/i) = P(X \in R_i/\pi_i) = \int_{R_i} f_i(X)dx \qquad (3)$$

Si existieran probabilidades a priori para estas dos poblaciones, la probabilidad de clasificación correcta o incorrecta, se calcula como el producto de la probabilidad condicional por la probabilidad a priori respectiva.

P(clasificación correcta como π_1) = $P(1/1)p_1$
P(clasificación incorrecta como π_1) = $P(1/2)p_2$
P(clasificación correcta como π_2) = $P(2/2)p_2$
P(clasificación incorrecta como π_2) = $P(2/1)p_1$

También se puede calcular la probabilidad total esperada de clasificación errónea: $PEM=P(1/2)p_2 + P(2/1)p_1$

Si $\{P(i/i)\}$ son las probabilidades de asignación correcta de una regla de clasificación d, y $\{P'(i/i)\}$ son las correspondientes a otra regla de clasificación d´; se dice que d es tan buena como d´, si se cumple que $P(i/i) \geq P'(i/i)$, para todo i=1,2,...,g. Además, d será mejor que d´, si al menos una de las desigualdades es

estricta. Si d es una regla para la cual no existe una mejor, entonces se dice que es admirable.

Otro aspecto de tomar en cuenta dentro del problema de construcción de reglas de clasificación, es el costo de hacer una clasificación equivocada, permitiendo de esta manera, asignar diferentes niveles de importancia a los diferentes tipos de error; por ejemplo, en diagnóstico médico se considera más dañino, en términos de sobrevivencia de un paciente, fallar en diagnosticar una enfermedad potencialmente fatal, que diagnosticar dicha enfermedad cuando en realidad no lo es. Si se define como C(i/j), el costo de clasificar un individuo en la población π_i cuando en realidad pertenece a la población π_j, para cualquier regla de clasificación se puede definir el costo esperado de clasificación se puede definir el costo esperado de clasificación errónea (CEM); por ejemplo, para el caso de dos poblaciones $CEM=C(2/1)P(2/1)p_1 + C(1/2)P(1/2)p_2$. Una “buena” regla de clasificación debe tener un CEM tan pequeño como sea posible.

II.2.6.1. EVALUACION DE LAS FUNCIONES DE CLASIFICACION

Todo procedimiento de clasificación tiene una posibilidad de clasificar erróneamente un determinado individuo, la cual es medida mediante la probabilidad de clasificación errónea; dicha probabilidad se construye en un índice importante para medir el comportamiento de un determinado criterio de clasificación o con medida de referencia para comparar varios procedimientos.

Cuando se conocen las distribuciones poblacionales completamente, el calculo de las probabilidades de clasificación errónea se hace relativamente fácil.

Algunas veces se conocen las distribuciones pero no los parámetros, en cuyo caso se tienen que estimar dichos parámetros usando los datos disponibles. De esta manera, tanto las regiones de clasificación como las probabilidades de clasificación errónea son también estimaciones.

Desde el punto de vista práctico, en gran parte de los problemas reales donde se aplica el **Análisis Discriminante**, no se conocen las formas de las distribuciones de las diferentes poblaciones; para estos casos se han desarrollado una serie de técnicas empíricas, basadas en los datos muéstrales disponibles, que son muy útiles en la estimación de las probabilidades de clasificación errónea. A continuación, se comentan el Método R (Resubstitución) siendo una de las técnicas que aparecen evaluadas en (Lachenbruch and Mickey, 1968).

A. Método R (Resubstitución)

Los datos usados para la derivación de las funciones discriminantes, son rehusados para la evaluación de las mismas. Por ejemplo,

$$\hat{P}(i/j) = \frac{n_i}{n_j} \qquad (4)$$

es un estimador de la probabilidad de clasificar erróneamente un individuo en la población π_i cuando en realidad pertenece a la población π_j, donde:

n_j = Cantidad de elementos en la muestra provenientes de π_j.

n_{ij} = Cantidad de elementos en la muestra provenientes de π_j, que según la función discriminante evaluada, quedan clasificados erróneamente como pertenecientes a π_i.

$$\sum_{i=1}^{g} n_{ij} = n_j \qquad (5)$$

Debido al hecho de usar las mismas observaciones para derivar las funciones discriminantes y luego para evaluarlas, el método resulta muy optimista en la estimación de las probabilidades de clasificación errónea; sin embargo, teniendo a mejorar su comportamiento a medida que los tamaños muéstrales se hacen bastante grandes.

PROPORCION DE CLASIFICACION ERRONEA

Otro criterio que se utiliza para la selección de variables, es la proporción de individuos mal clasificados; esto es, el criterio esta hecho con el propósito de seleccionar aquel subconjunto de variables que produzca la menor proporción de clasificación errónea (véase Murray, 1977, p. 246).

II.2.6.2. SELECCIÓN DE VARIABLES

Con frecuencia, investigadores que usan el **Análisis Discriminante** como técnica estadística para resolver un determinado problema, donde los datos disponibles son las respuestas a p variables de n individuos divididos en g grupos, desean hacer un análisis exploratorio previo al **Análisis Discriminante**, para seleccionar aquel subconjunto de variables que mejor discrimine entre los grupos. Algunas de las p variables originales, pueden ser ignoradas para los efectos del **Análisis Discriminante**; ya sea porque existen variables con poco poder discriminativo, debido a que los centroides de los grupos difieren muy poco con respecto a esas variables; o que existen variables redundantes, en el sentido de que dos o más variables comportan la misma información discriminante. La inclusión en el análisis de este tipo de variables, aparte de complicar el análisis, pueden llegar a desmejorar el comportamiento de las funciones discriminante, haciendo que se incrementen las probabilidades de clasificación errónea.

Uno de los procedimientos más usados para la selección de variables en el **Análisis Discriminante**, es el ***Procedimiento paso a paso ("Stepwise")***, similar al usado en el Análisis de Regresión Múltiple para seleccionar subconjuntos de variables predictoras. Dicho procedimiento, puede ser ascendente, descendente o mixto; y esta asociado a un determinado criterio de selección, basado generalmente en medidas de discriminación o proporciones de clasificación errónea.

El procedimiento paso a paso ascendente, comienza por seleccionar aquella variable individual que, de acuerdo al criterio de selección, aparece como la mejor; es decir, aquella variable individual que mejor discrimina entre los grupos. Luego, se combina la primera variable seleccionada, con cada una de las variables restantes (una a la vez), hasta obtener, de acuerdo con el criterio de selección, la mejor pareja; esto es, la segunda variable seleccionada es aquella variable que, combinada con la primera seleccionada, discrimina mejor entre los

grupos. De esta manera, tomando como base las dos primeras variables seleccionadas y siguiendo un procedimiento similar, se encuentra la mejor terna de variables. El proceso continua, cada vez incorporando una nueva variable al conjunto, hasta que de acuerdo con el criterio de selección, no se puede incorporar ninguna otra variable al conjunto, puesto que las variables restantes no incrementan significativamente el poder discriminativo de las variables ya seleccionadas.

El procedimiento paso a paso descendente, a diferencia del ascendente, comienza con todas las variables, luego de revisar las variables una a una y de acuerdo a determinado criterio, se descarta la peor, es decir, se descarta aquella variable que no esta aportando poder discriminativo adicional al resto de las variables. El proceso continúa descartando una variable en cada paso, hasta que no sea posible descartar más variables, ya que de acuerdo al criterio fijado, todas las variables que permanecen en el conjunto, son consideradas importantes en la discriminación.

Dentro de la ***Selección paso a paso*** de variables, posiblemente el procedimiento más eficiente es el ***mixto***, que consiste en combinar el ascendente y el descendente. Básicamente consiste de una ***selección ascendente***, pero antes de realizar el paso siguiente, se revisan una a una todas las variables previamente seleccionadas, para ver si es posible descartar una de ellas; cualquier variable descartada en un determinado paso, puede volver a ser seleccionada en un paso futuro. El procedimiento continúa hasta que no sea posible ni seleccionar ni descartar más variables. La inclusión de una determinada variable en un paso particular, se debe a que en ese momento, esa variable era la más importante; sin embargo, a medida que en los pasos siguientes entren otras variables al conjunto, la variable que previamente fue importante, ahora puede ser redundante por compartir la información discriminante con otras variables que entraron posteriormente, y en consecuencia, esta variable se convierte en candidata a ser removida.

El procedimiento paso a paso selecciona, de acuerdo al criterio utilizado, un conjunto óptimo de variables discriminantes; sin embargo, este conjunto no puede ser el mejor, ya que seleccionar el mejor conjunto, implicaría probar todas las posibles combinaciones de n variables para n = 1, 2, ... , p, y esto sería muy costoso en lo que a tiempo de computación se refiere.

Cada procedimiento de selección está asociado a un criterio de selección, el cual a su vez involucra un valor crítico, que al compararlo con ciertas medidas discriminativas propias del criterio, es el responsable de incluir o remover variables. A continuación se describe uno de los criterios más usados (véase Klecka, 1980, p. 54).

LAMBDA DE WILKS (F para incluir – F para remover)

La ***Lambda de Wilks***, toma en consideración tanto la dispersión entre grupos como la dispersión dentro de los grupos. En la selección de variables, es usada de la misma manera a como se usa en el análisis multivariante de varianza de una vía; es decir, se toma aquella variable, que al agregarla al conjunto de variables ya seleccionadas, produce la mayor separación significativa entre los grupos (menor valor significativo de lambda). De la misma manera, se remueve aquella variable, que al separarla del conjunto, la separación entre grupos con las variables restantes, no disminuye significativamente. En sustitución de lambda, se puede usar el estadístico **F (*basado en lambda*)**, utilizado para la prueba de diferencias entre vectores de medias de los diferentes grupos (véase Chatfield and Collins, p. 140). También, incluso recomendable, es usar la **F** para incluir (**F** parcial) en el caso de selección o la **F** para remover (**F** parcial) en el caso de estar descartando variables.

La **F** para incluir, sirve para probar la discriminación adicional que aporta la variable que está siendo considerada, después de tomar en cuenta la discriminación contenida en las otras variables ya seleccionadas; si este

incremento es el más significativo (**F** significativa más grande), la variable será seleccionada. El mínimo **F** para entrar (valor crítico), debe ser fijado de antemano, algunas veces es práctico usar el valor por defecto $\mathbf{F}_{1,\ \infty;0.95} = 4$. Los grados de libertad para la prueba, están dados por g – 1 para el numerador y n-q-g+1 para el denominador, donde q es el número de variables seleccionadas (incluyendo la que se está considerando) (véase Afifi and Azen, 1979, p. 310).

La **F** para remover, sirve para probar la disminución en discriminación que una determinada variable produciría, si es removida del grupo de variables ya seleccionadas; si ésta disminución es la menor entre las no significativas (**F** no significativa menor), la variable será removida. Para este caso los grados de libertad correspondientes son (g-1) y (n-q-g), y el valor crítico por defecto se puede fijar igual a 3.9, ligeramente menor que el mínimo valor para entrar.

Una vez seleccionadas las variables, se pueden ordenar de acuerdo al poder discriminativo de cada una de ellas, usando la **F** para remover. Así, la variable con la mayor **F** para remover, se considera como la que posee la mayor contribución en la discriminación, cuando actúa conjuntamente con las otras variables del grupo. Esta contribución, no es la misma que cuando las variables actúan solas; el poder discriminante de una variable cuando no interactúa con otras variables, puede ser visto en el primer paso del procedimiento.

Análisis Discriminante y el Análisis de Componentes Principales :

Si se compara la forma de la obtención de las ***Funciones Discriminantes*** con los ***Componentes Principales***:

Tabla II.5. Análisis de Componentes Principales vs. Análisis Discriminante.

Análisis de Componentes Principales	**Análisis Discriminante**
Son mutuamente ortogonales.	A pesar de estar incorrelacionadas entre sí, no son mutuamente ortogonales. Esto es, los ejes que representan las **Funciones Discriminantes**, no son un subconjunto de ejes obtenidos mediante una rotación rígida del sistema rectangular original de p variables (se conservan los ángulos), sino que son obtenidos por una rotación oblicua (ángulos entre los ejes del sistema no son rectos; véase Green, 1976, p. 104).
*Las **Funciones Discriminantes**, aparte de reducir la dimensionalidad, tal como lo hace los **Componentes Principales**, también son susceptibles a interpretación.*	

(**2, Dillon, William R.; Goldstein, Matthew)**

CAPITULO III

ESTUDIO COMPARATIVO ENTRE LAS ESPECIES DE INSECTOS: RHODNIUS ROBUSTUS Y RHODNIUS PROLIXUS (INSECTO TRANSMISOR DEL MAL DE CHAGAS)

III.1. INTRODUCCION

En base a las dificultades de clasificación de **Rhodnius prolixus** y **Rhodnius robustus**, se propone la comparación de dos muestras alopátrica de cada especie mantenidas, por varios años, en condiciones de laboratorio. De cada colonia se seleccionarán al azar 30 ejemplares machos y 30 ejemplares hembras, a los cuales se les estimarán 25 parámetros biométricos. Sus resultados serán revisados mediante el empleo de estadística descriptiva y algunas pruebas de Análisis Multivariante, en el próximo Capítulo.

III.2. RESEÑA HISTÓRICA

R. prolixus y **R. robustus** son especies estrechamente relacionadas bajo varios puntos de vista: morfológico, ecológico y más recientemente, isoenzimático. La conclusión general es que estas poblaciones están estrechamente relacionadas bajo un punto de vista reproductor y genético, más estudios deben hacerse para decidir el estado taxonómico definitivo de ambas especies.

Este pareciera ser el caso del género **Rhodnius**, ya que NEIVA y PINTO (1923) y LENT (1948) plantearon la homogeneidad morfológica de este género y las dificultades para un diagnóstico específico. Sin embargo, LENT y JURBERG (1969) al hacer una revisión del género, encuentran diferencias en la morfología de las genitalias externas, fundamentalmente en los machos de **Rhodnius prolixus** y **Rhodnius robustus**. En este sentido, LENT y WYGODZINSKY (1979)

aceptan a **R. Prolixus** y **R. Robustus** como las especies más parecidas del género, señalando que deberían ser las especies de más íntima relación filogenética, esta deducción la hacen en base a la morfología de las especies.

Por otra parte, **R. prolixus** fue considerado como triatomino estrictamente domiciliario hasta cuando GAMBOA (1961) demostró la presencia de poblaciones silvestres en el Municipio Ortiz del Estado Guárico. Posteriormente, GAMBOA (1970) refiere que el estudio comparativo de poblaciones de **R. prolixus** extra e intradomiciliarias arrojó diferencias morfólogicas y en los patrones de coloración. Además, indica el autor que los cruces realizados en el Instituto Venezolano de Investigaciones Científicas, entre poblaciones de uno y otro habitat demostraron fertilidad y en un evento científico, realizado en Caracas, en 1968 ambas poblaciones fueron reconocidas como pertenecientes a una sola especie.

El primer hallazgo de **R. robustus** en Venezuela lo efectúan LENT y VALDERRAMA (1973) en una palmera **Attalea maracaibensis** en el actual Municipio Zea del Estado Mérida, Venezuela. OTERO Y COLABORADORES (1975) lo señalan para los Estados Cojedes, Falcón y Táchira. Un año más tarde, TONN Y COLABORADORES (1976) lo encuentran en Apure, Barinas, Bolívar, Monagas, Sucre, Yaracuy y Zulia, con capturas en palmeras de género **Acrocomia, Mauritia** y sobre todo, **Scheelea**, en el que capturan 112 del total de 171 ejemplares de los diversos habitats, 13 de 18 palmeras de este último género (72.2%) resultaron positivas para este triatomino. Luego, ROSSELL (1977) lo refiere para el Estado Trujillo en la palmera **Acrocomia sclerocarpa**.

Posteriormente, BARATA (1981) determina que entre los huevos de las especies del género **Rhodnius**, los de **R. prolixus** y **R. robustus** por un lado y **R. nasutus** y **R. neglectus** por otro, son los más parecidos entre sí.

La caracterización especifica de las especies de triatominos se ha basado fundamentalmente en caracteres morfológicos, ello ha permitido el reconocimiento

de la mayoría de las especies. Sin embargo, en algunos casos, el diagnóstico específico se hace difícil debido a la semejanza de algunos grupos de triatominos. (Zeledón, 1983).

El género **Rhodnius** Stal, 1859 esta conformado por 13 especies. De ellas, RAMÍREZ PÉREZ (1987), reseña que 6 han sido halladas en territorio venezolano, entre las cuales se encuentran **Rhodnius prolixus** STAL, 1859 **y Rhodnius robustus** LARROUSE (1927).

Por otra parte, estudios ecofisiológicos realizados por ROSSELL (1984) muestran diferencias significativas entre una población de **R. prolixus** y una población de **R. robustus** en relación a la asignación energética. Sin embargo, estos resultados parecen diferir con los hallazgos de HARRY Y COLABORADORES (1992), quienes encuentran una elevada variabilidad en los perfiles isoenzimáticos de **R. prolixus**, los cuales no mostraron diferencias en relación al perfil isoenzimático de una muestra de **R. robustus**.

Otro resultado que llama la atención, más reciemtemente, en relación a **R. prolixus** y **R. robustus** resulta de la demostración de la posibilidad de flujo genético entre ambas especies GALINDEZ Y COLABORADORES (1994, a). Además, GALINDEZ Y COLABORADORES (1994, b) en un estudio morfológico, caracterizan poblaciones de una y otra especie pero no logran separar a **R. prolixus** de **R. robustus**, quizás debido a la elevada variabilidad de los caracteres biométricos considerados.

Desde el punto de vista ecoepidemiológico se tiene que ambas especies triatominas son transmisoras del agente etiológico de la **ENFERMEDAD DE CHAGAS**. En el medio silvestre pueden ser halladas en los mismos habitats, fundamentalmente en palmeras, lo cual permitiría el posible flujo génetico tanto naturalmente como por influencia del ser humano al utilizar algunos elementos de las palmeras para la fabricación de enseres o los propios techos de los hogares.

Lo señalado anteriormente sugiere la necesidad de la cuantificación precisa de caracteres de diversas poblaciones para su caracterización y su posterior comparación con poblaciones pertenecientes a otras especies como una posibilidad de contribuir al esclarecimiento taxonómico de **R. prolixus** y **R. robustus**. **(GALINDEZ G., 1994)**

III.3. MATERIALES Y METODOS

Para la caracterización sistemática de **Rhodnius prolixus** y **Rhodnius robustus** se realizó un estudio morfométrico, empleando triatominos de las colonias mantenidas en el insectario “Pablo Anduze” del Centro de Investigaciones Trujillanas “José Wiltremundo Torrealba” del Núcleo Universitario “Rafale Rangel”, de la Universidad de los Andes.

Las especies y procedencias se identificaron de la siguiente manera:

1.- Rhodnius prolixus: R.P.

Dos ejemplares:

RPBR : R. prolixus de Venezuela

RPCO : R. prolixus de Colombia

111 : RPBR hembras

112 : RPBR machos

121 : RPCO hembras

122 : RPCO machos

2.- Rhodnius robustus: R.R.

Dos ejemplares:

RRCA : R. robustus de Venezuela

RRCO : R. robustus de Colombia

211 : RRCA hembras

212 : RRCA machos

221 : RRCO hembras

222 : RRCO machos

Mantenimiento de las colonias:

Las colonias de triatominos estudiadas fueron mantenidas en envases de vidrio, con una capacidad promedio de 4 litros, cubiertos con tela de organdí, sujetas con bandas de goma. Dentro de los mismos se colocó un trozo de papel plegado dispuesto verticalmente para aumentar la superficie útil para los insectos y permitir su desplazamiento hacia la fuente alimentaria al momento de la ingesta.

Los envases conteniendo los insectos fueron mantenidos en un cuarto a 26±2°C y 70±10% de humedad relativa, con período de luz y oscuridad de 12 horas e ingesta sanguínea sobre gallina quincenalmente durante 30 minutos en cada ocasión.

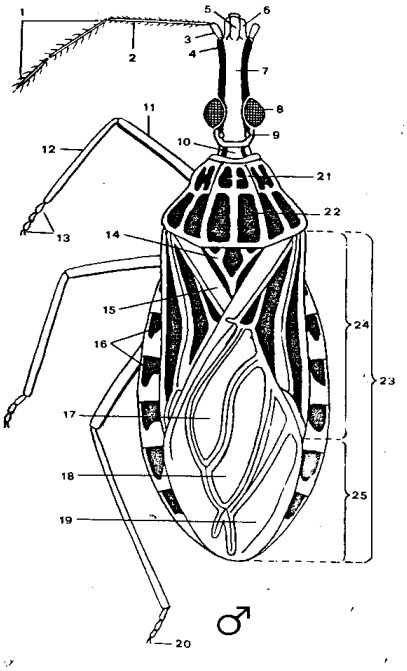
III.4. ESTUDIO MORFOMÉTRICO

La muestra estudiada fue de 60 individuos adultos (30 hembras y 30 machos) para cada una de las colonias estudiadas, siendo observados en un microscopio estereoscópico Nikon con un aumento de 30X.

Los datos almacenados de los ejemplares estudiados se pueden ver en el ***Anexo A***.

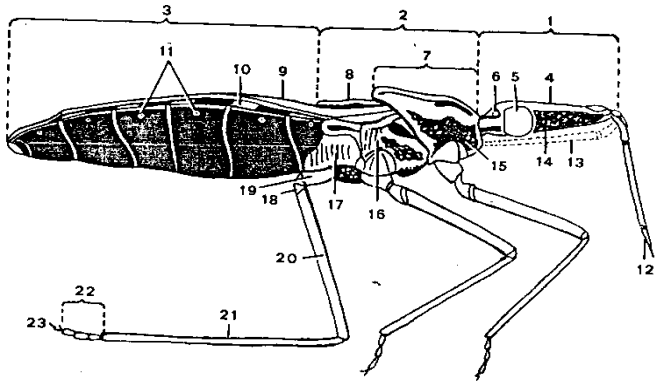
Los parámetros seleccionados para el estudio, las ilustraciones que permiten identificar al insecto así como la distribución geográfica del mismo en Venezuela, se pueden ver en las páginas siguientes.

Ver Tabla de Excel Archivo Características



Vista dorsal del macho adulto de Rhodnius Prolixus.

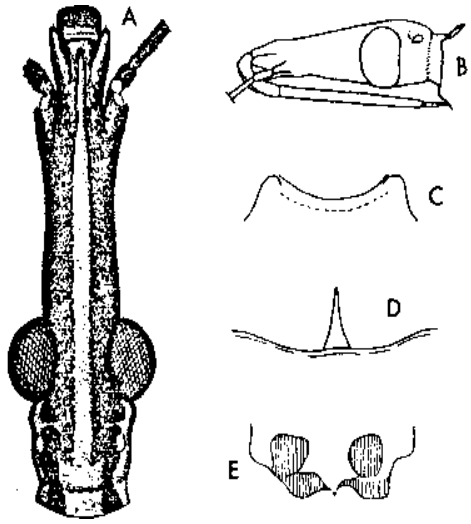
1. Flagelo, 2. Pedicelo, 3. Escapo, 4. Tubérculo antenífero, 5. Postclípeo, 6. Lámina maxilar, 7. Frente, 8. Ojo compuesto, 9. Ocelo (ojo simple), 10. Cuello, 11. Fémur, 12. Tibia, 13. Tarso, 14. Mesoscutelo, 15. Clavo, 16. Conexivo, 17. Célula anal, 18. Célula cubita, 19. Célula media, 20. Uñas, 21. Lóbulo anterior del pronoto, 22. Lóbulo posterior del pronoto, 23. Ala anterior, 24. Región corácea, 25. Región membranosa.



Vista lateral de la hembra adulta de Rhodnius Prolixus

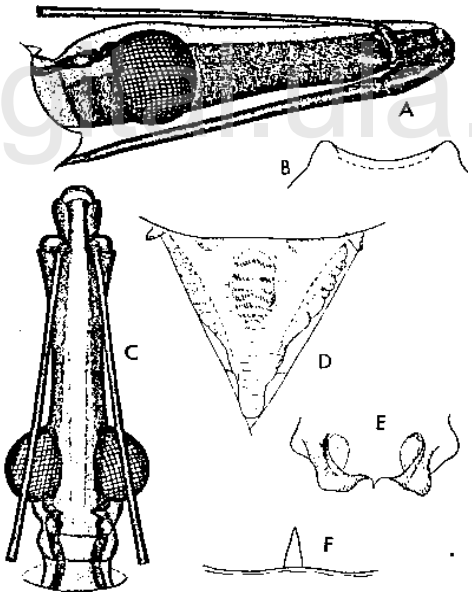
1. Cabeza, 2. Tórax, 3. Abdomen, 4. Frente, 5. Ojo, 6. Ocelo, 6. Pronoto, 8. Mesoscutelo, 9. Ala anterior, 10. Conexivo, 10. Estigmas respiratorios, 12. Trompa replegada, 13. Labio en reposo, 14. Gena, 15. Propleura, 16. Mesopleura, 17. Mesopleura, 18. Trocánter, 19. Metacoxa, 20. Fémur, 21. Tibia, 22. Tarso, 23. Uñas.
Tomado de Ramírez Pérez, 1.969.

Fig. 6. Insecto Rhodnius Prolixus



Rhodnius Prolixus

A. Cabeza anterior, B. Cabeza, vista lateral, C. Cuello,
D. Proceso del medio de pygophore, E. Pavoneos del plato básales



Rhodnius Robustus

A. Cabeza, vista lateral, B. Cuello, C. Cabeza, aspecto dorsal,
D. Escutelo esquemático, E. Pavoneos del plato básales,
F. Proceso posterior de pygophore.

Fig. 7. Insecto Rhodnius Prolixus vs Rhodnius Robustus
Ver Tabla de Excel Archivo Distribución en el pais.

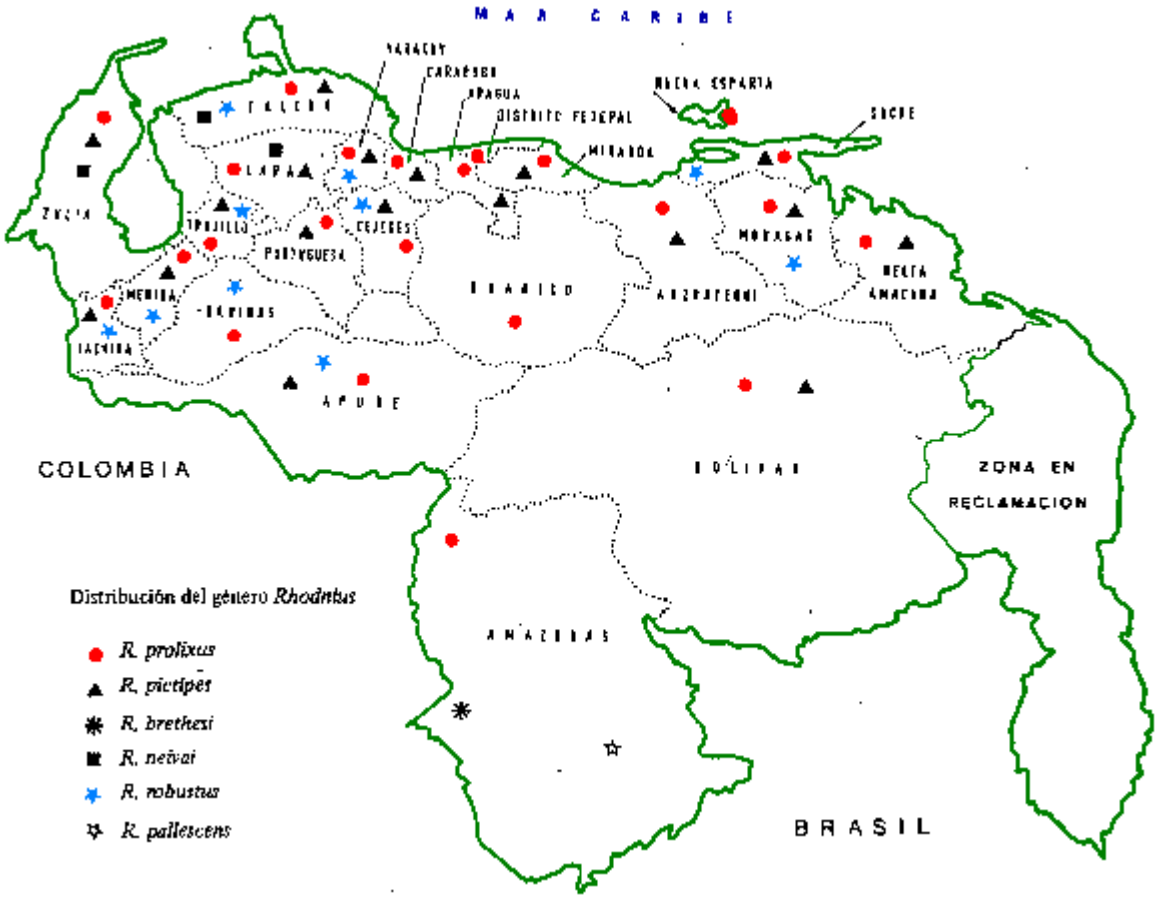


Fig. 8. Mapa de Venezuela con la distribución geográfica del **Género Rhodnius.**

III.5. ENFERMEDAD DE CHAGAS

Definición:

Es la infección de mamíferos y de triatominos producida por el ***Trypanosoma cruzi***. En el hombre, la infección puede ser congénita o adquirida y

afecta, en grado variable, diversos órganos y sistemas, especialmente el corazón y el tubo digestivo.

Biología:

El ***Trypanosoma cruzi*** es un protozoo mastigóforo perteneciente a la familia ***Trypanosomatidae***, en cuyo ciclo biológico intervienen mamíferos y un insecto vector. Los huéspedes mamíferos pueden ser el hombre y algunos animales, domésticos (el perro o el gato) o silvestres (diversos mamíferos, especialmente, los roedores y los carnívoros).

Epidemiología:

La severidad e irreversibilidad de las lesiones cardíacas y de otros órganos, provocan invalidez y mortalidad entre los grupos económicamente activos. Sin embargo, las estadísticas sanitarias no reflejan la verdadera magnitud del problema, porque la enfermedad prevalece en zonas suburbanas o rurales, donde la atención médica no capta, en su integridad, la importancia de la infección.

El conocimiento de la magnitud de la infección chagásica y su repercusión sobre la salud y la economía de los países latinoamericanos, varía grandemente y, en especial, sus formas clínicas.

En Venezuela, el área endémica abarca el 80% del territorio y comprende más de 4 millones de personas expuestas al riesgo de infectarse. El principal vector es el R. prolixus. El compromiso cardíaco, medido por estudios electrocardiográficos y clínicos, contituye un importante problema médico asistencial y epidemiológico. No se han descripto megas digestivos en este país.

En Venezuela, la cardiopatía es muy importante, pero no se han descrito magaformaciones.

Patología:

En la enfermedad de Chagas, existe compromiso de los órganos ricos en sistema, linfoidomacrofágico (ganglios linfáticos, hígado y bazo), sistema nervioso central, miocardio y órganos huecos, especialmente el tubo digestivo. En la fase aguda de la infección, se observa un aumento del volumen de los ganglios, espleno y hepatomegalia, meningoencefalitis y cardionegalia por la dilatación de las cavidades del corazón. En la fase crónica, el compromiso se centra fundamentalmente en el miocardio y en el tubo digestivo. **(ATIAS, NEGHME, 1979)**

CAPITULO IV

APLICACIÓN DEL ANALISIS MULTIVARIANTE A DOS ESPECIES DE INSECTOS:

RHODNIUS PROLIXUS Y RHODNIUS ROBUSTUS

IV.1. ANALISIS DE RESULTADOS

Los valores obtenidos en el estudio morfométrico fueron procesados con el paquete estadístico **S.P.S.S.** versión 9.0. (Ver ***Anexo***)

Presentando de cada método las salidas que representaron información relevante para el caso estudiado, se ordenaron como sigue:

IV.1.1. Análisis Factorial
IV.1.2. Análisis de Cluster
IV.1.3. Análisis Discriminante,

se pudo obtener los siguientes resultados:

Factor Analysis

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	11.145	50.660	50.660	11.145	50.660	50.660	5.273	23.970	23.970
2	2.233	10.151	60.811	2.233	10.151	60.811	5.071	23.052	47.021
3	1.460	6.638	67.449	1.460	6.638	67.449	4.494	20.428	67.449
4	.890	4.047	71.496						
5	.761	3.459	74.955						
6	.657	2.985	77.940						
7	.635	2.888	80.828						
8	.533	2.423	83.252						
9	.487	2.213	85.465						
10	.443	2.016	87.481						
11	.415	1.888	89.369						
12	.362	1.647	91.015						
13	.333	1.512	92.528						
14	.329	1.495	94.023						
15	.302	1.373	95.396						
16	.246	1.117	96.513						
17	.221	1.004	97.518						
18	.189	.861	98.378						
19	.134	.608	98.986						
20	.120	.547	99.533						
21	.102	.462	99.995						
22	1.155E-03	5.250E-03	100.000						

Extraction Method: Principal Component Analysis.

Tabla IV.1. Criterio o Regla de Kaise

Esta Tabla muestra la extracción de los factores, siendo la representación de las variables originales (Características del insecto) con una pérdida mínima de información.

En este caso se han extraído 3 factores, ya que, tenían Autovalores mayores o iguales a 1.0; indicando la proporción de la varianza explicada por un factor en una variables particular.

Juntos ellos consideran el 67.449% de variabilidad de los datos originales es decir, los factores explican los datos en un 67.449%%.

Ver Gráfico en Statgraphics Archivo Insecto Folio Factorial

Gráfico 1. Criterio Scree_Test de Castell

Esta gráfica muestra los *Autovalores* para cada uno de los 22 factores.

Los *Autovalores* son proporcionales al porcentaje de la variabilidad en los datos atribuibles a los factores.

Se puede ver una línea horizontal en 1.0, valor usado para decidir en extraer los 3 factores.

Ver Tabla de Excel Archivo Resultados Factorial Hoja1

Tabla IV. 2. Matriz Factorial y Comunalidades Estimadas.

La **Tabla IV. 2** muestra las ecuaciones que estiman los factores comunes. ***Por ejemplo****, el 1er. Factor común tiene la ecuación:*

$\mathbf{F_1}$ = 0.706*__C1__ + 0.768*__C2__ + 0.700*__C3__ + 0.307*__C4__ + 0.639*__C5__ + 0.424*__C6__ + 0.869*__C7__ + 0.728*__C8__ + 0.650*__C9__ + 0.600*__C10__ + 0.714*__C11__ + 0.767*__C12__ + 0.728*__C13__ + 0.708*__C14__ + 0.841*__C15__ + 0.439*__C16__ + 0.878*__C17__ + 0.877*__C18__ + 0.873*__C19__ + 0.742*__C20__ + 0.660*__C21__ + 0.710*__C22__

A esta matriz se le llama también ***Matriz Factorial o de Cargas***. Cada carga significa la Correlación que existe entre cada variable en cada factor.

Los valores más llamativos son los más grandes para cada factor, bien sean positivos o negativos, es decir, una Correlación en el mismo sentido o en sentido opuesto, se puede verificar en la ***Matriz de Correlación***.

Por ejemplo:

En el factor 2 las variables **C5** y **C6** están correlacionadas en sentido opuesto, es decir, entre más pequeña sea la distancia entre los ojos mayor será el diámetro de los ojos, de los insectos en estudio.

La última columna nos muestra las *Comunalidades Estimadas*, interpretándose como la proporción de la varianza en cada variable en base a los 3 factores. Siendo ésta la suma de las cargas al cuadrado.

Por ejemplo:

Para **C1** sería:

$$(0.71)^2 + (0.33)^2 + (-0.21)^2 = 0.65$$

En las **Gráficas 2** y **3** se muestra en los Cuatro Cuadrantes la distribución de los individuos en estudio y las características para cada Componente según la *Matriz de Cargas*.

Donde podemos concluir que los individuos que se encuentren en el **II** y **III Cuadrante** tienen sus características pequeñas. Mientras que los que se encuentran en los **Cuadrantes I** y **IV** según su posición tienen sus características de mayor o menor tamaño.

Por ejemplo:

- En el **Cuadrante I** se encuentran aquellos individuos que tienen entre otras, las siguientes características grandes:
  C1: Longitud del insecto,
  C6: Diámetro del ojo,
  C7: Amplitud máxima de la cabeza en los ojos (en vista dorsal),
  C16: Amplitud del pronoto en los ángulos anteroculares,...

- En el **IV Cuadrante** el individuo **207** (**2/2/1**: Insecto Rhodnius Robustus – Colombiano- Femenino) tiene la distancia postocular (**C4**) y la distancia interocular (vista dorsal) (**C5**) grandes mientras que el diámetro del ojo (**C6**) y su longitud o talla (**C1**) las tiene pequeñas.

Y con respecto a los individuos que se encuantran cerca del cero se puede decir que son los *Individuos Promedios*, es decir, que tienen sus características según sus promedios mostrados en la **Tabla IV.4**.

Ver Gráfico en Statgraphics Archivo Insecto Folio Componentes

Gráfico 2. Biplot del Componente 1 vs. el Componente 2.

Ver Gráfico en Statgraphics Archivo Insecto2 Folio Componentes

Gráfico 3. Biplot del Componente 2 vs. el Componente 3.

Ver Tabla de Excel Archivo Resultados Factorial Hoja2

Tabla IV.4. Valores promedios de los insectos según sus características.

Los Individuos Promedios tendrían los valores anteriores en cada una de sus características.

Por ejemplo:

El individuo **144** (**2/1/1**: Rhodnius Robustus-Venezolano-Femenino) es un Individuo Promedio, es decir, sus características tienen valores promedios.

Component Transformation Matrix

Component	1	2	3
1	.625	.593	.508
2	-.187	-.518	.835
3	-.758	.617	.213

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalizatior

Tabla IV.5. Matriz de Componentes de transformación

Esta Tabla muestra los valores óptimos que permiten, la rotación de los factores. Se escogió el método de rotación más usado Varimax para obtener dicha matriz.

Rotated Component Matrix[a]

	Component		
	1	2	3
C1	.539	.119	.588
C2	.413	.474	.451
C3	.372	.497	.341
C4	-9.83E-02	.714	-.109
C5	.384	.779	-.123
C6	8.369E-02	-.116	.866
C7	.370	.495	.679
C8	.441	.605	.183
C9	.103	.685	.354
C10	.891	.169	-.113
C11	.454	.666	6.950E-02
C12	.468	.594	.240
C13	.493	.382	.379
C14	.329	.578	.315
C15	.579	.379	.500
C16	-5.47E-02	7.877E-02	.840
C17	.616	.291	.630
C18	.528	.505	.488
C19	.720	.449	.308
C20	.904	.139	.185
C21	.450	.280	.419
C22	.253	.531	.466

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Tabla IV. 6. Matriz con componentes rotados

Esta tabla muestra las ecuaciones que estiman los factores comunes después de realizar una rotación.

Este procedimiento es adicional, se realiza en caso de que no sea fácil poder apreciar los factores, las características y los individuos que se están estudiando. Esta Matriz se interpreta y se representa gráficamente igual que la Matriz de Factorial o de Cargas. Se obtiene de multiplicar la Matriz Factorial por la Matriz de Componentes de Transformación.

Ver Tabla de S.P.P.S Archivo Cluster

Tabla IV.7. Combinación de Cluster para los Insectos.

La **Tabla IV.7**. muestra como fueron combinados los individuos según la distancia que existe entre los Cluster, en los pasos indicadas.

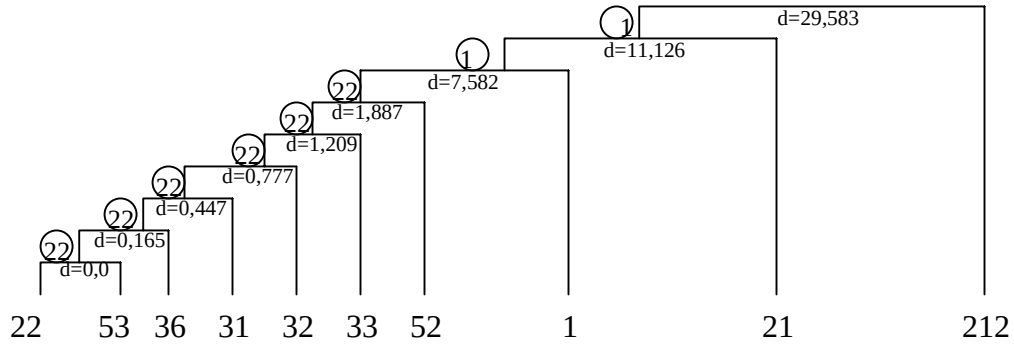
Las columnas indican: el paso en que ocurrió la reunión de los individuos, la combinación de los individuos según al Cluster que pertenezcan, la distancia entre los individuos, el paso del Cluster donde apareció primero y el próximo paso en donde se combinarán otros individuos.

Por ejemplo:

Según la primera fila, los individuos se fueron uniendo así:

En el Paso 1 : **22** (1/1/1) con **53** (1/1/2) a una distancia 0,000
" " " 44: el Cluster **22** " **36** (1/1/2) " " " 0,165
" " " 143: " " **22** " **31** (1/1/2) " " " 0,447
" " " 194: " " **22** " **32** (1/1/2) " " " 0,777
" " " 216: " " **22** " **33** (1/1/2) " " " 1,209
" " " 229: " " **22** " **52** (1/1/2) " " " 1,887
" " " 237: " " **22** " **1** (1/1/1) " " " 7,582
" " " 238: " " **1** " **21** (1/1/1) " " " 11,126
" " " 239: " " **1** " **212(**2/2/2) " " " 29,583

El Dendrograma de este ejemplo sería así:



Se puede apreciar que se combinaron las hembras Rhodnius Prolixus de Venezuela con los machos y finalmente a esta primera combinación se une un macho Rhodnius Robustus de Colombia.

Ver Tabla de S.P.P.S Archivo Clustersolovariables

Tabla IV.8. Combinación de Cluster para las características.

La **Tabla IV.8.** muestra como fueron combinadas las características según la distancia que existe entre los Cluster en los pasos indicados.

Pro ejemplo:

En la primera fila se observa las siguientes uniones:

En el paso 1: **C6** (Diamétro del ojo) con **C22** (Grosor del fémur) a una distancia de 1,409.

En el paso 3: **C6** con **C5** (Distancia inter-ocular, distancia más corta entre los ojos) a una distancia de 2,222.

En el paso 5: **C5** con **C4** (Distancia post-ocular) a 3,634.

En el paso 9: **C4** con **C10** (Distancia inter-ocular, mínima distancia entre los ocelos) a una distancia de 7,972.

En el paso 10: **C4** con **C9** (Longitud del ojo, en vista lateral) a 54,563.

En el paso 17: **C4** con **C7** (Amplitud máxima de la cabeza en los ojos, en vista dorsal) a 171,774.

En el paso 20: **C4** con **C2** (Longitud de la cabeza) a 1568,291.

En el paso 21: **C2** con **C1** (Talla o longitud del insecto, en vista dorsal) a una distancia de 57.062,03.

Ver Tabla de S.P.P.S Archivo Clustersolovariables

Gráfica 4. Dendrograma combinaciones de los Cluster para las características.

Discriminant

(Según la Especie)

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.874[a]	100.0	100.0	.683

a. First 1 canonical discriminant functions were used in the analysis.

Tabla IV. 9. Poder discriminante para la Función Discriminante para la Especie.

Esta Tabla indica que los grupos tienen una asociación de 0,683 por tanto, esta función tiene poder discriminante.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.534	146.937	8	.000

Tabla IV.10. Nivel de significación entre las especies.

Este resultado revela un nivel de significación menor que 0,05 por ello, se concluye que existe diferencia significativa entre las especies.

La Función Discriminante tiene poder significativo con un nivel de confianza del 95%.

Analysis

Box´s Test of Equality of Covariance Matrices Of Canonical Discriminant Functions

Test Results

Box's M		2.088
F	Approx.	2.079
	df1	1
	df2	169932.0
	Sig.	.149

Tests null hypothesis of equal population covariance matrices of canonical discriminant functions.

Tabla IV. 11. Prueba de igualdad de matrices de covarianza.

Para este Análisis es necesario saber si los datos cumplen con el supuesto de igualdad entre las matrices de covarianza. Y dado los resultados obtenidos indican que se aceptan la hipótesis nula (las matrices de covarianza son iguales), ya que, el nivel de significación es mayor que 0,05.

Stepwise Statistics

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
8	C13	.536	43.874	.635
	C18	.296	36.881	.619
	C20	.413	22.453	.586
	C5	.543	8.358	.553
	C3	.534	9.904	.557
	SEXO	.645	11.778	.561
	C4	.727	7.623	.551
	C19	.241	5.247	.546

Tabla IV.12. Variables incluidas en el Análisis Discriminante para las especies.

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
8	C1	.478	.239	.919	.532
	C2	.498	.240	2.516	.528
	C6	.469	.241	.012	.534
	C7	.370	.241	.000	.534
	C8	.495	.240	1.826	.529
	C9	.570	.230	.659	.532
	C10	.344	.231	.942	.532
	C11	.403	.240	1.531	.530
	C12	.452	.239	1.574	.530
	C14	.536	.238	.040	.534
	C15	.346	.236	1.667	.530
	C16	.741	.241	.020	.534
	C17	.247	.231	2.707	.527
	C21	.589	.240	2.843	.527
	C22	.517	.241	1.432	.530

Tabla IV. 13. Variables no incluidas en el Análisis Discriminante para las especies.

Al usar el Método de paso a paso se obtuvo 8 características de los insectos que discriminan más.

Para ello el Método hace las siguientes consideraciones:

En cada paso entra una variable que minimiza el Lambda de Wilks. (Se presenta el paso 8 el cual contiene a las variables involucradas).

El valor de F parcial mínimo para entrar tiene que ser 3.84 y un máximo F parcial para remover la variable de 2.71.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
SEXO	.402
C3	.406
C4	-.307
C5	-.371
C13	.799
C18	-.998
C19	.444
C20	.678

Tabla IV. 14. Coeficientes estandarizados de la Función Discriminante para las especies.

Este resultado muestra los coeficientes estandarizados de la Función Discriminante:

Y = 0,402*SEXO + 0,406*C3 – 0,307*C4 – 0,371*C5 + 0,799*C13 – 0,998*C18 + 0,444*C19 + 0,678*C20

Se puede apreciar las características que más contribuyen dentro de la Función Discriminante, según el valor relativo del coeficiente son:

C18 (Amplitud del esculeto),
C13 (Longitud del segundo segmento del rostro (RII) y
C20 (Amplitud máxima del abdomen).

Functions at Group Centroids

ESPECIE	Function 1
1.00	-.931
2.00	.931

Unstandardized canonical discriminant functions evaluated at group means

Tabla IV. 15. Centroides para las especies según la Función Discriminante.

Esta Tabla muestra el valor del Centroide según la Función Discriminante para cada Especie. Cuyos valores permiten imaginar que los insectos se encuentran agrupados en diferentes sectores del eje de coordenadas.

Calassification Statistics

Classification Results[a]

		ESPECIE	Predicted Group Membership 1.00	2.00	Total
Original	Count	1.00	103	17	120
		2.00	27	93	120
	%	1.00	85.8	14.2	100.0
		2.00	22.5	77.5	100.0

a. 81.7% of original grouped cases correctly classified.

Tabla IV. 16. Predicción al clasificar los grupos de especies.

Esta Tabla proporciona la estimación de las probabilidades de clasificación errónea. Así de los 120 individuos pertenecientes a la Especie 1, el 85,8% fueron bien clasificados y el 14,2% debieron estar en la Especie 2. De los 120 de la Especie 2, el 77,5% fueron clasificados correctamente mientras que el 22,5% restantes fueron clasificados como pertenecientes a la Especie 1.

Por tanto,

$\hat{P}(1/2) = 0{,}142$	$\hat{P}(2/1) = 0{,}225$

Separate_Groups Graghs

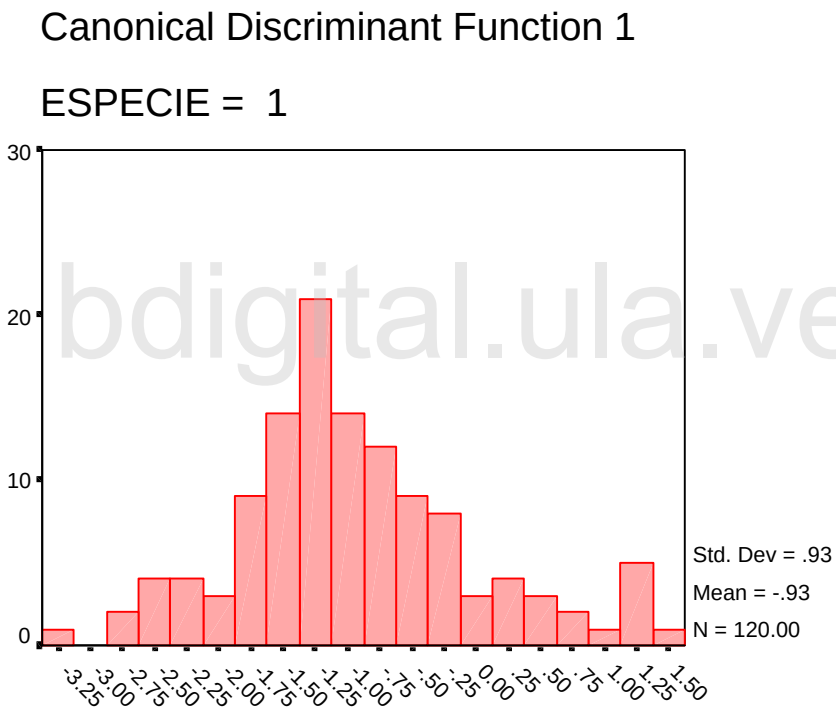


Gráfico 5. Distribución de los datos para la especie Rhodnius Prolixus.

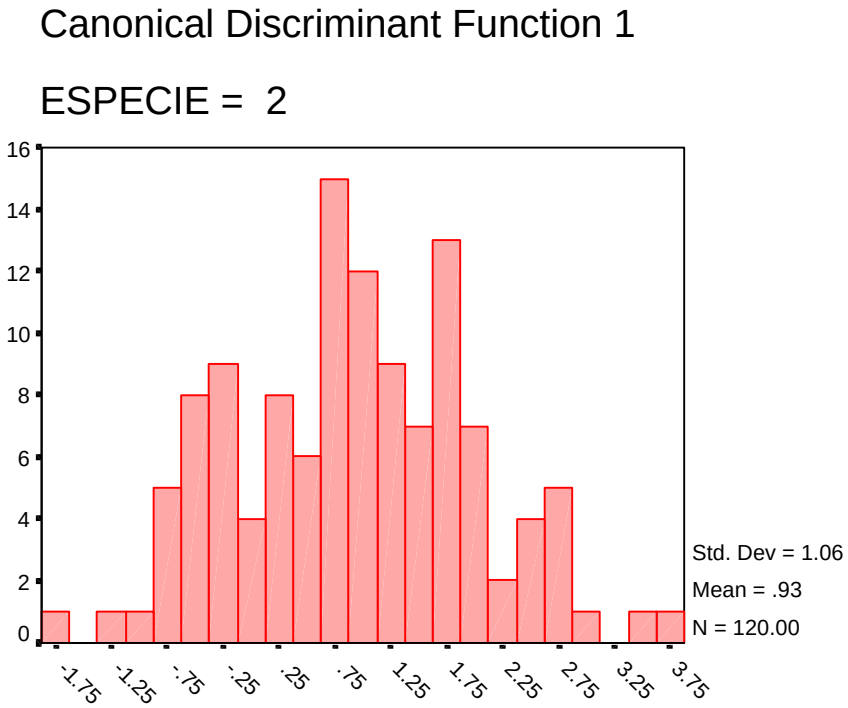


Gráfico 6. Distribución de los datos para la especie Rhodnius Robustus.

Discriminant

(Según la Población)

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2.095[a]	100.0	100.0	.823

a. First 1 canonical discriminant functions were used in tl analysis.

Tabla IV. 17. Poder discriminante para la Función Discriminante para la Población.

Esta Tabla indica que los grupos tienen una asociación de 0,823 por tanto, esta función tiene poder discriminante.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.394	217.059	10	.000

Tabla IV.18. Nivel de significación entre las especies.

Este resultado revela un nivel de significación menor que 0,05 por ello, se concluye que existe diferencia significativa entre las especies.

La Función Discriminante tiene poder significativo con un nivel de confianza del 95%.

Stepwise Statistics

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
12	C2	.592	23.860	.357
	C11	.602	11.793	.340
	C10	.270	26.708	.361
	C20	.216	11.608	.340
	SEXO	.506	51.877	.397
	C18	.339	7.854	.334
	C13	.459	13.356	.342
	C9	.669	11.182	.339
	C17	.238	15.166	.345
	C6	.606	9.939	.337
	C3	.499	7.353	.334
	C22	.668	5.816	.331

Tabla IV.19. Variables incluidas en el Análisis Discriminante para las poblaciones.

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
12	C1	.405	.203	.079	.323
	C4	.810	.215	2.723	.319
	C5	.297	.216	.107	.323
	C7	.191	.191	.055	.323
	C8	.548	.216	.318	.323
	C12	.502	.216	.487	.322
	C14	.649	.216	1.400	.321
	C15	.367	.211	1.074	.322
	C16	.480	.216	.011	.323
	C19	.249	.213	.400	.322
	C21	.573	.215	2.477	.320

Tabla IV. 20. Variables no incluidas en el Análisis Discriminante para las poblaciones.

Al usar el Método de paso a paso se obtuvo 12 características de los insectos que discriminan más.

Para este caso:

El valor de F parcial mínimo para entrar tiene que ser 3.84 y un máximo F parcial para remover la variable de 2.71.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
SEXO	.737
C2	.487
C3	-.305
C6	-.320
C9	.322
C10	-.758
C11	.348
C13	-.423
C17	.623
C18	.382
C20	.577
C22	.235

Tabla IV. 21. Coeficientes estandarizados de la Función Discriminante para las poblaciones.

Este resultado muestra los coeficientes estandarizados de la Función Discriminante:

Y = 0,737*SEXO + 0,487*C2 – 0,305*C3 – 0,320*C6 + 0,322*C9 – 0,758*C10 + 0,348*C11 - 0,423*C13 + 0,623*C17 + 0,382*C18 + 0,577*C20 + 0,235*C22

Se puede apreciar las características que más contribuyen dentro de la Función Discriminante, según el valor relativo del coeficiente son:

C10 (Distancia interocular mínima distancia entre los ocelos),
SEXO y
C20 (Amplitud del pronoto en los ángulos humerales).

Functions at Group Centroids

	Function
POBLACIO	1
1.00	-1.441
2.00	1.441

Unstandardized canonical discriminant functions evaluated at group means

Tabla IV. 22. Centroides para las poblaciones según la Función Discriminante.

Esta Tabla muestra el valor del Centroide según la Función Discriminante para cada Pobalción. Cuyos valores permiten imaginar que los insectos se encuentran agrupados en diferentes sectores del eje de coordenadas.

Calassification Statistics

Classification Results[a]

			Predicted Group Membership		
		POBLACIO	1.00	2.00	Total
Original	Count	1.00	113	7	120
		2.00	9	111	120
	%	1.00	94.2	5.8	100.0
		2.00	7.5	92.5	100.0

a. 93.3% of original grouped cases correctly classified.

Tabla IV. 23. Predicción al clasificar los grupos de población.

Esta Tabla proporciona la estimación de las probabilidades de clasificación errónea. Así de los 120 individuos pertenecientes a la Población 1, el 94,2% fueron bien clasificados y el 5,8% debieron estar en la Población 2. De los 120 de la Población 2, el 92,5% fueron clasificados correctamente mientras que el 7,5% restantes fueron clasificados como pertenecientes a la Población 1.

Por tanto,

$\hat{}$ $\hat{}$

P(1/2) = 0,580 P(2/1) = 0,750

Separate_Groups Graghs

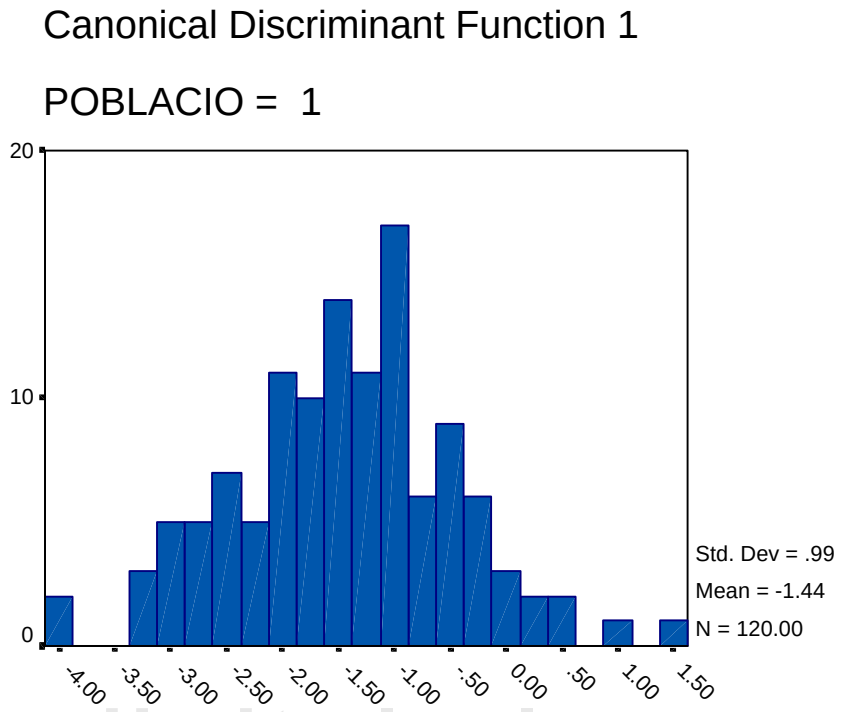


Gráfico 7. Distribución de los datos para la población Venezolana.

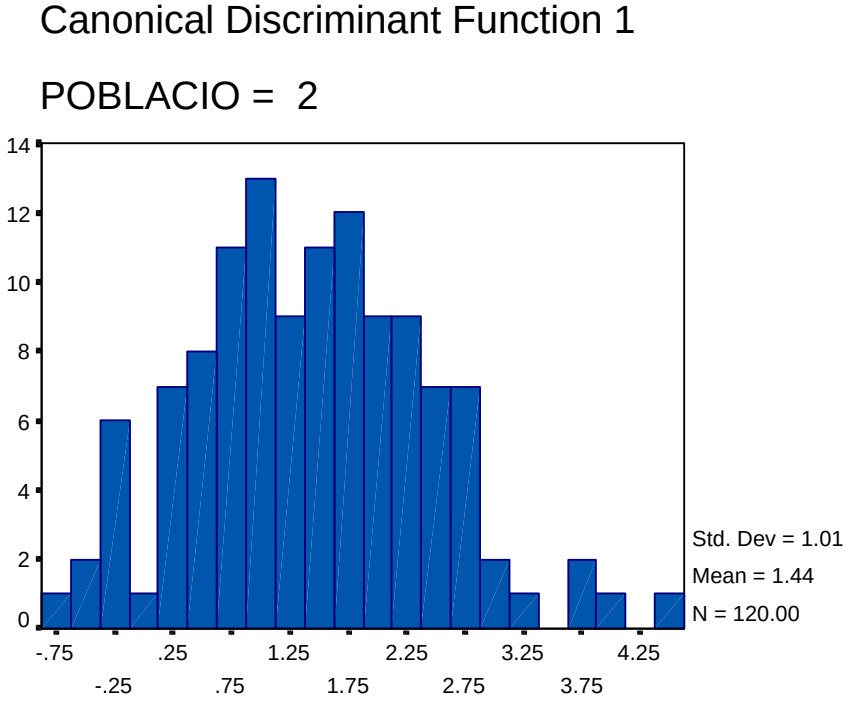


Gráfico 8. Distribución de los datos para la población Colombiana.

CONCLUSIONES

Una vez finalizado este proyecto y alcanzados los objetivos propuestos se pudo comprobar que es factible la aplicación de las técnicas del Análisis Multivariante mediante la herramienta computacional.

Se pudo revisar una serie de conocimientos relacionados con el Análisis Multivariante, siendo sintetizados para presentar así, las ideas principales sobre los métodos y sus aplicaciones. De esta manera, ofrecer un soporte al análisis de los resultados.

Con respecto al estudio comparativo entre los insectos de la Especie Rhodnius se obtuvo una información general sobre la enfermedad del Mal de Chagas y antecedentes de estudios comparativos entre estas especies.

Después de aplicarse algunos de los Métodos del Análisis Multivariante, se puede concluir:

Análisis Factorial:

- Reduce el número de variables en Factores lo que permite interpretar los datos en base a ellos.
- Se puede apreciar cuán correlacionadas están las variables.
- Con la ayuda gráfica se complementa el análisis, en él se aprecian los individuos en estudio y las variables medidas, la relación entre ellos y como se distribuyen según los factores extraídos.
- Los resultados arrojados por éste análisis permite establecer como son las características de los insectos sin necesidad de ir a las tablas donde están plasmadas estas medidas.

- Según la Especie, la Población y el Sexo de estos insectos se pudo apreciar como se agrupan según sus características, es decir, la tendencia que ellos presentan.

Análisis de Cluster:

- Se pudo observar la forma cómo se van agrupando los insectos por la similitud de sus características, en otras palabras, se detalla que individuos conforman un “Cluster”.
- Este análisis complementa el Análisis Factorial, en una presentación más detallada.

Análisis Discriminante:

- Con este análisis se puede concluir que éstos insectos son diferentes según la Especie, la Población y el Sexo, ya que, sus características lo establecen así. Esta diferencia es significativa con un nivel de confianza del 95%.
- A través de la Función Discriminante se pudo apreciar qué características contribuyen en determinar a qué grupo pertenece un insecto según la Especie y la Población, es decir, nos indica las características que hay que tomar en cuenta a la hora de hacer las mediciones a estos insectos, ahorrando con ello tiempo y dinero. Esa función discriminante también permitirá predecir a que grupo pertenece un insecto cualquiera al sustituir los valores medidos de cada una de sus características.

Finalmente, el paquete estadístico S.P.S.S versión 9.0 para Windows es una herramienta sencilla y muy útil, por su ambiente amigable a través de menús y cajas de diálogo, porque permite obtener resultados en forma eficiente y por la gran cantidad de herramientas estadísticas que tiene.

REFERENCIAS BIBLIOGRAFICAS

1. **AFIFI, A. A., AND AZEN S. P.** *Statistical Analysis: A computer Oriented Approach* (2nd. Ed.). New York: Academic Press, 1979.
2. **ANDERSON, T. W**. *An Introduction to Multivariate Statistical Analysis*. New York: Chapman-Hall, 1958.
3. **ATIAS, A., NEGHME, A.** *Parasitología Clínica.* Editorial Intermedica. Buenos Aires, Argentina. 1979. Págs. 215-229.
4. **BARATA, J. M.** (1948). *Aspectos morfólogicos de ovos de Triatominae. II.- Cracterísticas macroscópicas e exocoriasis de dez espécies do Gênero* ***Rhodnius*** *Stal, 1859 (Hemiptera-Reduviidae). Rev. Saúde Públi. S. Paulo 15: 490-542.*
5. **BARATA, J. M., SANTOS, J.L. Y LEITEE, C.A.P.** (1980) *Aspectos morfológicos de ovovs de Triatominae. I.- Mensuraçâo de dez espéciees do Gênero* ***Rhodnius*** *Stal, 1859 (Hemiptera-Reduviidae). Rev. Brasil. Entomol. 24(3/4): 197-214.*
6. **BATISTA FOGUET, JOAN MANUEL y MARIA DEL ROSARIO MARTINEZ ARIAS**. *Análisis Multivariante: Análisis de Componentes Principales*. Barcelona:Editorial Hispano Europea, S.A., 1989.
7. **BISQUERA ALZINA, R.** (1989), *"Introduccion Conceptual al Analisis Multivariable: Un enfoque Informatico con los Paquetes S.P.S.S – X, BMDP, LISREL Y SPAD"* Vol. 1, PPU, S.A."
   [URL: www.uam.es/estructura/facultades/Economi.../obtencion.html]
   Last modified 17-Mar-97 - page size 7K - in Spanish [Translate]:
8. **CARCAVALLO, R. U., TONN, R.J., ORTEGA, R. BETANCOURT, P. y CARRASQUERO, B.,** 1978. *Notas sobre la biología, ecología y distribución geográfica de* ***Rhodnius prolixus*** *Stal, 1859 (Hemiptera, Reduviidae, Triatominae).Bol. Dir. Malariol. San. Amb. 18(3):175-198.*
9. **CHATFIELD, C., AND COLLINS, A. J.** *Introduction to Multivariate Analysis.* London: Chapman and Hall, 1980.

10. **COCHRAN, W. G. COMMENTARY** on *"Estimation of Error Rates in Discriminant Analysis"*. Technometrics, 1968, 10 (1), 204-205.

11. **COOLEY, W. W., LOHNES, P. R.** *Multivariate Data Analysis.* New York: John Wiley, 1971.

12. **DILLON, WILLIAM R.; GOLDSTEIN, MATTHEW** *"Multivariate Analysis Methods and applications".* Wiley Series in Probability and Mathematical Statistics.

13. **DIXON, W. J.** *Biomedical Computer Programs*. Berkeley: University of California Press, 1974.

14. **EVERITT, B. S.** *Graphical Techniques for Multivariate Data*. London: Heinemann Educational Books, 1978.

15. **GALINDEZ GIRON, ITAMAR.** Trujillo, Octubre 1989. *Revision critica de la estructura y composición del genero Rhodnius Stal, 1859 (Hemiptera, Reduviidae, Triatominae).* Centro de enfermedades tropicales. J. W. Torrealba

16. **GALINDEZ GIRON, ITAMAR.** Trujillo, Enero 1994. *Entomología y Vectores.* Comité Científico Editorial 1994-1995.

17. **GAMBOA, J. C.** 1961. *Comprobación de* ***Rhodnius prolixus*** *extradoméstico. Bol. Inf. Dir. Malario. San. Amb. 1(5): 139-142.*

18. **GAMBOA, J. C.** 1970. *La población silvestre de* ***Rhodnius prolixus*** *en Venezuela. Bol. Inf. Dir. Malario. San. Amb. 10(5-6): 186-207.*

19. **GNANADESIKAN, R.** *Methods for Statistical Data Analysis of Multivaratee Observations.* New York: John Wiley, 1977.

20. **GRAYBILL, F. A.** *Theory and Application of the Linear Model.* Belmont, California: Duxbury Press, 1976.

21. **GREEN, P. E.,** *Analyzing Multivariate Data.* Hinsdale, IL: Dryden, 1972.

22. **GREEN, P. E., AND CARROL, J. D.** *Mathematical Tools for Applied Multivariate Analysis*. New York: Academic Press, 1976.

23. **HARRY, M., GALINDEZ, I. Y CARIOU, M. L.** 1992. *Isozyme variability and differentiation between* ***Rhodnius prolixus, Rhodnius robustus*** *and* ***Rhodnius pictipes,*** *vectors of Chagas disease in Venezuela. Med. Veter. Entomol. 6: 37-43.*

24. HARRIS, R. J. *A primer of Multivariate Statistica.* New York: Academic Press, 1975.

25. HOTELLING, H. *Analysis of a complex of statistical variables into principal components*, J. Educ. Psychol., 24:417-441 y 498-520 (1933).

26. JOHNSON, R. A., AND WICHERN D. W. *Applied Multivariate Statistical Analysis.* Englewood Cliff, New Jersey: Prentice-Hall, 1982.

27. JOLLIFFE, I. T. *Discarding Variables in a Principal Component Analysis. I: artificial data.* Applied Statistics, 1972, 21, 160-173.

28. JOLLIFFE, I. T. *Discarding Variables in a Principal Component Analysis. II: real data.* Applied Statistics, 1973, 22, 21-31.

29. KLECKA, W. R. *Discriminant Analysis*. Beverly Hills: Sage Publications, 1980.

30. LACHENBRUCH, P. A., AND MICKEY, M. R. *Estimation of Error Rates in Discriminant Anlysis.* Technometrics, 1968, 10 (1), 1-11.

31. LENT, H. 1948. *O Género **Rhodnius** Stal, 1859 (Hemiptera, Reduviidae). Rev. Brasil. Biol. 8(3): 297-339.*

32. LENT, H. y VALDERRAMA, A. 1973. *Hallazgos en Venezuela del triatomino **Rhodnius robustus** Larrousse, 1927 en la palma **Attalea maracaibensis** Martius (Hemiptera, Reduviidae). Bol. Inf. Dir. Malariol. San. Amb. 13(5/6): 175-179.*

33. MARDIA, K. V., KENT, J. T., AND BIBBY, J. M. *Multivariate Analysis.* London: Academic Press, 1979.

34. MORRISON, D. F. *Multivariate Statistical* Methods (2nd. ed.). Tokyo: McGraw-Hill Kogakusha, 1967.

35. MURRAY, G. D. *A Cautionary Note on Selection of Variables in Discriminant Analysis.* Applied Statistics, 1977, 26 (3), 246-250.

36. PEARSON, K. *On lines and planes of closed fit to system of pont in space,* Phil. Mag., 6:559-572 (1901)

37. PLA, Laura E. *"Análisis Multivariado: Método de Componentes Principales"* Universidad Nacional Experimental -Francisco de Miranda - Coro, Flacón,

VENEZUELA. Secretaria General de la Organización de los Estados Americanos.

38. RAMIREZ PEREZ, J. 1987. *Revisión de los triatominos (Hemiptera, Reduviidae) en Venezuela.* Bol. Dir. Maralio. San. Amb. 15(5):217-230.

39. ROSSELL, O., MOGOLLON, J. Y PACHECO, J.E. 1977. *Presencia de **Rhodnius robustus** Larrousse, 1927 (Hemiptera, Reduviidae) en el Estado Trujillo, Venezuela.* Bol. Dir. Maralio. San. Amb. 17(3):230-233.

40. SNEDECOR, GEORGE W. y WILLIAM G. COCHRAN, 2da. Imp. *Métodos estadísticos*. México: Compañía Editorial Continental, 1974.

41. TONN, R.J., CARCAVALLO, R. U. y ORTEGA, R. 1976. *Notas sobre la biología, ecología y distribución geográfica de **Rhodnius robustus** Larrousse, 1927 (Hemíptera, Reduviidae).*Bol. Dir. Malariol. San. Amb. 16(2):158-162.

Consultas por INTERNET en AltaVista :

1. Introducción al Análisis Factorial
URL: ww3.uniovi.es/user_html/herrero/Dpto_Ps...r.1/indice.html]
Introducción al Análisis Factorial. Copyright Marcelino Cuesta y Fco.J.Herrero Dpto. Psicología Universidad de Oviedo...
Last modified 11-Nov-95 - page size 3K - in Spanish [Translate]

ANEXO A

Tablas de Datos

Ver Tabla de Excel Archivo Presenta Datos

ANEXO B

Paquete Estadístico S.P.S.S. versión 9.0

S.P.S.S. (Statistical Program System Software) es una herramienta computacional para resolver problemas estadísticos, permitiendo la manipulación, el análisis y representación de los datos en estudio.

La barra de menús contiene submenús para abrir archivos, hacer cálculos estadísticos y crear gráficos. También proporciona acceso a la mayoría de las características de S.P.S.S., desde la modificación de valores de datos al cambio de fuentes.

Los pasos básicos para el análisis de datos con S.P.S.S. resulta muy fácil. Todo lo que tiene que hacer es:

1) Introducir los datos a S.P.S.S.. Puede abrir un archivo de datos de S.P.S.S. guardado previamente; leer una hoja de cálculo, dBASE o un archivo de datos de texto; o bien introducir sus datos directamente en el Editor de datos.
2) Seleccionar un procedimiento en los menús para calcular estadísticos o crear gráficos. Puede seleccionar un procedimiento para realizar análisis estadísticos o seleccionar un procedimiento para crear gráficos de alta resolución.
3) Seleccionar las variables que desea utilizar en el análisis. Las variables del archivo de datos aparecen en un cuadro de diálogo para el procedimiento.
4) Examinar los resultados.

De esta manera sencilla se trabaja con el S.P.S.S. siendo entonces una herramienta que permite en forma rápida y veraz el análisis de datos y por ende la toma de decisiones. **(Bisquera , 1989)**

Ver PowerPoint Archivo Anexo B