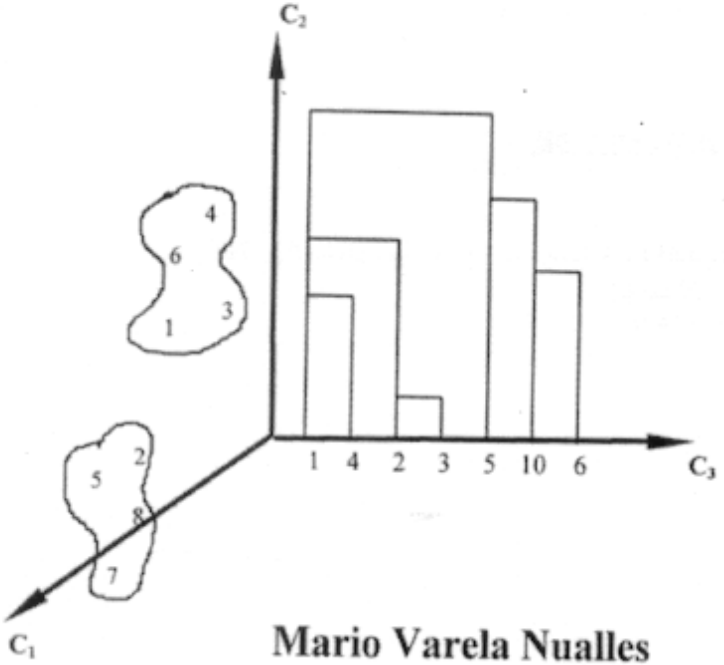
INCA

Análisis Multivariado de Datos. Aplicación a las Ciencias Agrícolas



Mario Varela Nualles

Matemática Aplicada

La Habana, 1998

Corrección y edición. **María Mariana Pérez Jorge**
Diseño y realización: **Yamila Isabel Díaz Bravo**

SOBRE LA PRESENTE EDICIÓN:

© Instituto Nacional de Ciencias Agrícolas (INCA), 1998
© Mario Varela Nualles
ISBN: 959-7023-04-0

Ediciones INCA
Gaveta postal 1, San José de las Lajas,
La Habana, Cuba, CP 32 700

Introducción

En una investigación agrícola constantemente están interactuando factores que pueden ir desde las características fisiológicas de una planta, propiedades del suelo, hasta factores de tipo ambientales o climáticos, los cuales ejercen un efecto conjunto en los resultados finales de un experimento. En estos casos, según plantean Gladys Linares, Liliam Acosta y Viviam Sistach (1986), no es adecuado llevar a cabo una serie de análisis estadísticos univariados para cada una de las variables, ya que en ellos se ignoran estructuras de correlaciones, llegando incluso en ocasiones a falsear los resultados finales de la investigación.

El Análisis Multivariado, cuyas primeras ideas surgieron a principios de siglo, es la parte de la Estadística Matemática que se encarga del análisis de datos correspondientes a mediciones múltiples, tomadas en un conjunto de individuos o elementos de la muestra (Cooley, 1971). Este autor cita a Kendal (1957), el cual plantea que el usuario de las técnicas multivariadas está interesado en las interrelaciones de p variables observadas en un conjunto de n individuos; considera las variables dependientes entre ellas mismas, por lo que no deben ser analizadas de forma independiente.

La aplicación de las técnicas del Análisis Multivariado, apoyado en los avances de la computación, ha aumentado su gama de acción a varias esferas del conocimiento, particularmente a problemas relacionados con las ciencias agrícolas, en donde cada vez son más los investigadores que hacen uso de estas herramientas de la Estadística Matemática para interpretar sus resultados con la mayor confiabilidad posible. Así, por ejemplo, María C. González (1991) plantea que los métodos del Análisis Multivariado constituyen una herramienta útil, tanto para evaluar la variabilidad fenotípica como para conocer la contribución relativa de distintos caracteres a la misma; por otra parte, Miriam Alvarez (1982), plantea que las técnicas del Análisis Multivariado permiten clasificar diferentes tratamientos mediante un grupo numeroso de variables, por lo que tienen un uso frecuente en la clasificación de variedades en diferentes cultivos.

A pesar de la aplicación práctica e inmediata que tienen estas técnicas estadísticas, en ocasiones se piensa que resulta suficiente saber que determinado método de análisis multivariado es utilizado con determinado fin, compatible con nuestros propósitos, para ya con ello estar seguros de la efectividad de su uso, sin hacer un estudio previo de los datos. Este proceder trae consigo que se arriben a conclusiones completamente erróneas sobre un problema determinado. En tal caso, no es que el método haya sido ineficiente, sino que, por el contrario, estas técnicas fueron mal utilizadas por el investigador.

Para que un desarrollo agrícola sea técnicamente adecuado, para que ofrezca el menor número de despilfarros, pérdidas de tiempo y de esfuerzos, y para que resulte provechoso en el mayor grado posible, es necesario que se adquieran constantemente nuevos conocimientos respecto al mejor aprovechamiento de los campos, de los métodos más eficaces de cultivo en cada caso concreto, y de las variedades de planta y abonos que conviene utilizar. Esto requiere de un estudio previo que ha de estar a cargo de especialistas; es indispensable si se quiere que el desarrollo agrícola de nuestros países en vías de desarrollo, sea a la vez rápido y eficiente, que quienes en él intervienen como técnicos dispongan de un arsenal estadístico lo más completo posible (Panse y Sukhatme, 1967).

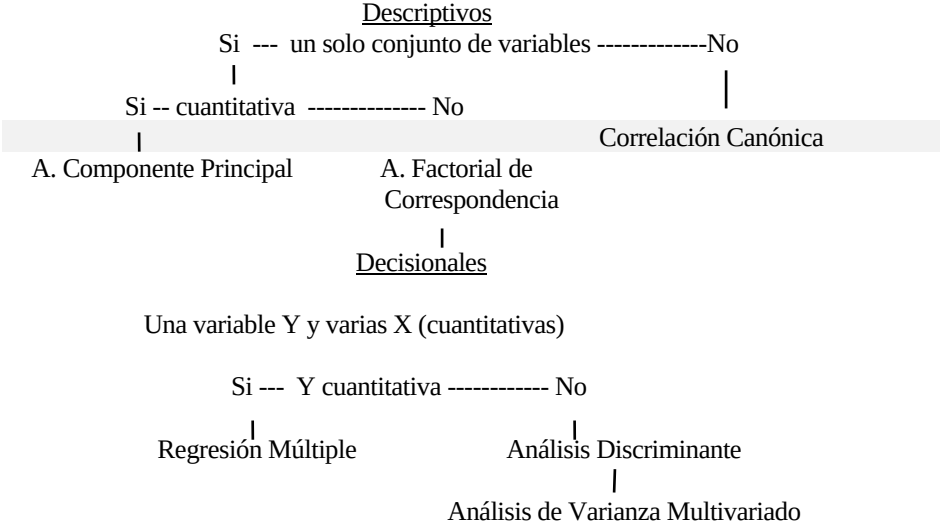
Debido a la importancia que tiene la aplicación de los métodos multivariados en el análisis de datos vinculados a problemas en la rama agrícola así como a la escasa bibliografía vinculada con el tema existente en nuestro idioma, consideramos que se hace necesario profundizar en el conocimiento de estas técnicas de análisis.

Desarrollo

♦ ***Clasificación de los métodos multivariados***

Los métodos multivariados se clasifican fundamentalmente atendiendo a los fines que se persiguen en la investigación : Gnanadesikan (1977) y Gladys Linares, Liliam Acosta y Viviam Sistach (1986) reagrupan en sus trabajos los técnicas multivariadas, siguiendo como criterio el propósito que se tenga en determinada investigación, el cual puede ser descriptivo o decisional.

Clasificacion de acuerdo con los fines que se persiguen



Estos mismos autores dan otra forma de clasificación de estos métodos, la cual combina con la ya mencionada, pero con la diferencia que centra su análisis en la estructura de la matriz inicial de datos.

Clasificación según matriz de datos

1- Matriz no particionada
- Análisis de Componentes Principales
- Análisis Factorial Clásico
- Análisis Factorial de Correspondencias

2- Matriz particionada *a priori*

por fila	por columna
- Análisis Discriminante	- Análisis Regresión
- Análisis Varianza Multivariado	- Correlaciones Canónicas

3- Se desea particionar por filas o columnas
- Cluster Análisis o métodos de clasificación

Por otra parte, Krzonowski (1988) da una clasificación, en la que también tiene en cuenta la estructura inicial de la matriz de datos; al respecto, este autor reagrupa las técnicas de Correlaciones Canónicas y Análisis Factorial Discriminante, como dos métodos en los que existen datos agrupados: en el primer caso, por variables y, en el segundo caso, por individuos. Como métodos multivariados en los que los datos no aparecen agrupados, este autor considera al Análisis de Componentes Principales y al Cluster Análisis.

De cualquier manera, todas estas formas de clasificación permiten en última instancia el reconocimiento de determinado método multivariado como una herramienta estadística capaz de dar cumplimiento a determinado objetivo propuesto; sin embargo, por muy bien que haya sido identificado el método, si hacemos un uso indiscriminado del mismo, los resultados finales pueden conducir a interpretaciones completamente sesgadas sobre un problema objeto de estudio.

Análisis de componentes principales

♦ *Introducción*

Una de las técnicas del Análisis Multivariado más difundida en la actualidad lo es, sin lugar a dudas, el Análisis de Componentes Principales, propuesto por Pearson (1901) y desarrollado por Hotelling (1933).

En la mayoría de los estudios exploratorios, los investigadores coleccionan observaciones sobre un gran número de variables, sin saber inicialmente a ciencia cierta cuáles son las más importantes o más útiles para un trabajo científico (Gladys Linares, Liliam Acosta y Viviam Sistach, 1986). De hecho, el investigador trata de incluir todas las variables que sospecha que puedan tener alguna conexión con el tema. Su próximo paso es reducir estos datos (trabajar con menos variables), para hacer menos engorroso los cálculos y facilitar la interpretación de los resultados experimentales.

El Análisis de Componentes Principales tiene como finalidad, construir un conjunto de nuevas variables o componentes, con la característica de que en este conjunto la mayor parte de la información o variabilidad inicial va a concentrarse en los primeros ejes o componentes. Este resultado permite a su vez reducir la dimensionalidad del problema, facilitando la caracterización de los elementos de la muestra y la búsqueda de estructuras de correlación entre variables.

Estas nuevas variables o componentes, no son más que combinaciones lineales de las variables originales. Se construyen de forma tal que entre ellas no haya correlación alguna; además, tienen la característica de que cada una presenta varianza máxima, es decir, explica la mayor cantidad posible de información inicial.

♦ *Nociones del fundamento matemático*

Como en todo método multivariado, se parte de la matriz inicial de datos X:

$$X=\begin{bmatrix} & \vdots & \\ \cdots & ij & \cdots \\ & \vdots & \end{bmatrix}_{nxp}$$

n: # de individuos
p: # de variables
n≥p

Así, el elemento *ij* de la matriz representa el valor observado de la variable *j* en el individuo *i*. En este caso, es oportuno señalar que las *p* variables deben ser de naturaleza continua, puesto que el método trabaja con el coeficiente de correlación de Pearson, diseñado para medir la relación lineal existente entre variables continuas.

A partir de esta matriz se estiman el vector de medias $u_{px1}=(u_1,u_2,...,u_p)$ y la matriz de varianzas y covarianzas E_{pxp} por medio de $Xmedia_{px1}$ y S_{pxp} respectivamente.

El objetivo es encontrar *p* funciones $(Y_1,Y_2,...,Y_p)$, que se expresan como combinación lineal de las variables originales, las cuales se denominan componentes principales.

Sean:

$$Y_1 = \sum_{j=1}^{p} A_{1j} X_j, \ldots, Y_p = \sum_{j=1}^{p} A_{pj} X_j$$

Así, para hallar Y_1, es decir, la primera componente, es necesario encontrar los coeficientes A_{1j} $j=1..p$, de forma tal que la varianza de Y_1 sea máxima, sujeta a la condición:

$$\sum_{j=1}^{p} A_{1j} = 1$$

lo cual asegura la unicidad de la solución.

Para hallar Y_2 (segunda componente), es necesario encontrar los coeficientes A_{2j} $j=1..p$, de forma tal que la covarianza de Y_2 con Y_1 sea igual a cero; además, debe cumplirse que la varianza de Y_2 sea máxima y que:

$$\sum_{j=1}^{p} A_{2j} = 10$$

Nótese que en el caso del cálculo de Y_2, se exige una condición más; es por ello, que debe obtenerse que la varianza de Y_2 va a ser menor o igual que la varianza de Y_1.

Para hallar Y_3 (tercera componente), es necesario encontrar los coeficientes A_{3j} $j=1..p$, de forma tal que la covarianza de Y_3 con Y_2 y la covarianza de Y_3 con Y_1 sean ambas iguales a cero. Además, Y_3 debe tener varianza máxima y los coeficientes deben cumplir la condición:

$$\sum_{j=1}^{p} A_{3j} = 10$$

Por la misma razón, la varianza de Y_3 debe ser menor o igual que la varianza de Y_2 y Y_1.

El resto de las componentes se calculan por el mismo algoritmo, hasta llegar a la componente p.

La solución a este problema se traduce en encontrar los valores y vectores propios de la matriz S de varianzas y covarianzas. (Hope, 1968 y Dempster, 1969). Así, por ejemplo, el mayor de los valores propios de S será el valor correspondiente a la varianza de Y_1, y su vector propio asociado se identifica con los coeficientes A_{1j}, $j=1..p$. El segundo valor propio (estableciendo un orden decreciente), será el valor correspondiente a la varianza de Y_2 y su vector propio asociado se representa por los coeficientes A_{2j} $j=1..p$; y así sucesivamente hasta llegar a Y_p.

Cooley (1971) y Gnanadesikan (1977) ofrecen en detalles la demostración de este resultado.

Nótese que finalmente las componentes cumplen la propiedad de estar incorrelacionadas y la varianza de la primera va a ser mayor que la de la segunda y así sucesivamente.

La suma de las varianzas de todas las componentes va a ser igual a la traza de la matriz S de varianzas y covarianzas, debido a que éstas varianzas no son más que los valores propios. Ahora bien, esta no es más que la suma de las varianzas de las variables originales X_i, $i=1..p$, ya que estos son los elementos de la diagonal S.

$Var(Y_1)+Var(Y_2)+...+Var(Y_p)=Traza(S)=Var(X_1)+Var(X_2)+...+Var(X_p)$

Es por este resultado, que el Análisis de Componentes Principales puede ser utilizado en la reducción de dimensionalidad, es decir, tratar de explicar con menos componentes la información inicial que se recoge en la matriz inicial de datos X.

Una vez construidas las componentes, se procede al cálculo de la correlación de cada componente con las variables iniciales; este es un paso muy importante, ya que a partir de estas correlaciones, es que se va a tener un criterio para caracterizar los ejes o componentes.

Es oportuno señalar que algunos autores plantean que, en ocasiones, resulta más efectivo en lugar de calcular los valores y vectores propios a partir de la matriz *S* de varianzas y covarianzas, calcularlos a partir de la matriz de correlaciones. Al respecto, Morrison (1979) plantea que se debe utilizar esta última cuando las variables originales están medidas en diferente escala, ya que a través del cálculo de la matriz de correlaciones se estandarizan los datos.

♦ *Ejemplos de aplicación*

Ejemplo 1: En un experimento de campo se estudió el comportamiento en el rendimiento de 10 variedades de calabaza, sometidas a ocho condiciones diferentes de estrés de temperatura y humedad. Como resultado se obtuvo la siguiente matriz de datos:

	E1	E2	E3	E4	E5	E6	E7	E8
1	8.44	8.01	2.50	3.29	1.33	0.80	1.84	1.19
2	9.57	5.63	3.60	6.27	0.37	1.35	1.76	1.54
3	10.93	5.42	3.70	6.10	0.90	0.99	1.30	0.14
4	8.60	8.87	5.10	4.90	0.90	0.91	1.22	1.57
5	9.60	3.37	1.45	2.80	0.35	0.56	0.97	0.62
X= 6	6.58	2.53	2.83	2.97	0.57	0.13	3.10	0.13
7	4.66	3.64	2.80	1.52	0.35	0.32	2.74	0.37
8	5.70	4.65	2.74	1.83	0.92	0.20	1.46	0.37
9	5.33	3.03	1.92	2.91	1.09	0.27	1.71	0.54
10	3.40	3.68	2.38	1.62	0.41	1.65	0.72	0.52

Así, el elemento *xij* de esta matriz corresponde al valor del rendimiento de la variedad de calabaza *i* en el ambiente o condición de estrés *Ej*.

En este caso, como las variables se miden en la misma escala, puesto que todas ellas corresponden a valores de rendimiento, se utilizó como matriz a diagonalizar para obtener las componentes principales, la matriz de varianzas y covarianzas.

Matriz de varianzas y covarianzas:

	E1	E2	E3	E4	E5	E6	E7	E8
E1	6.19	2.51	0.89	3.60	0.13	0.21	-0.33	0.44
E2		4.53	1.49	1.88	0.38	0.39	-0.43	0.86
E3			1.05	1.13	0.04	0.15	0.00	0.25
E4				3.07	0.07	0.36	-0.18	0.43
E5					0.12	-0.03	-0.01	0.02
E6						0.26	-0.22	0.13
E7							0.55	-0.10
E8								0.29

Los valores y vectores propios de esta matriz se ofrecen a continuación:
valores propios:
9.7472 2.8282 0.9187 0.5284 0.2059
contribución a la variación total:
67.3% 19.5% 6.3% 4.0% 1.4%
vectores propios (coeficientes para la combinación lineal):

	Y_1	Y_2	Y_3	Y_4	Y_5
E1	0.6963	0.5156	0.3921	0.1448	0.1827
E2	0.4882	-0.7725	0.2624	0.1057	-0.1023
E3	0.1976	-0.2493	-0.5561	0.1715	0.6572
E4	0.4784	0.1957	-0.6264	-0.2249	-0.3635
E5	0.0304	-0.0757	0.0951	0.1009	-0.1361
E6	0.0544	-0.0677	-0.1001	-0.4871	-0.1230
E7	-0.0539	0.0626	-0.2381	0.7904	-0.3954
E8	0.0952	-0.1509	0.0012	-0.1252	-0.4496

La contribución a la variación total de cada valor propio, expresa qué por ciento de la variación total explica la componente asociada. Este por ciento de contribución se obtiene de la siguiente forma:

El valor 19.5 % que es la contribución de la componente *Y2*, se obtiene a partir de buscar qué por ciento de 16.06 representa 2.8282 (valor propio asociado). En este caso, 16.06 es la traza de la matriz de varianzas y covarianzas, la cual es a su vez una magnitud de la variabilidad total.

Por otra parte, a partir de los vectores propios se puede obtener cada una de las componentes principales; así, por ejemplo:

Y1= 0.6963E1+0.4882E2+0.1976E3+0.4784E4+0.0304E5+0.0544E6-
0.0539E7+0.0952E8

Nótese que con las dos primeras componentes se explica un 86.8% de la variabilidad total (67.3% + 19.5%). Esto indica que en lugar de trabajar con las ocho variables iniciales, se puede a partir de ahora hacer cualquier tipo de inferencia de los datos a partir de las dos primeras componentes, teniendo siempre en cuenta que la primera componente en este caso es casi tres veces más importante que la segunda.

Comoquiera que las componentes son variables ficticias, se hace necesario darles un sentido biológico a partir de su relación con las variables iniciales. Esta relación se busca a partir de la correlación de las variables iniciales con las componentes.

Correlaciones entre las variables iniciales y las dos primeras componentes:

	Y1	*Y2*
E1	**0.9142**	0.3672
E2	**0.7541**	**-0.6428**
E3	**0.6325**	**0.4000**
E4	**0.8975**	0.1978
E5	0.2818	-0.3777
E6	0.3476	-0.2331
E7	-0.2384	0.1491
E8	0.5800	**-0.4953**

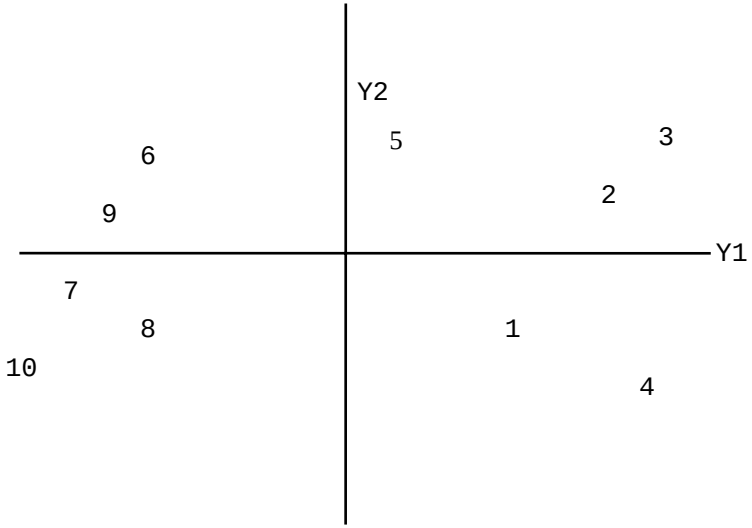
La correlación de las componentes con las variables iniciales indica cuáles variables caracterizan la componente. En este caso, se puede concluir que la componente 1 (la más importante) está caracterizada fundamentalmente por el comportamiento de las variedades en las cuatro primeras condiciones de estrés (E1, E2, E3 y E4); mientras que la componente 2, se caracteriza por un efecto aunque menos pronunciado de los ambientes E2, E3 y E8.

En el caso de la componente Y_1, como su correlación con E_1 ,E_2, E_3 y E_4 es positiva (directa), se puede afirmar que a medida que el valor de la componente 1 aumenta, se incrementa el rendimiento de las variedades de calabaza en las cuatro primeras condiciones de estrés.

Por otra parte, en el caso de la segunda componente, a medida que el valor de la misma se incrementa, los resultados en rendimiento para las variedades de calabaza en las condiciones de estrés E_2 y E_8 disminuyen (correlación negativa), no ocurriendo así con E_3, la cual tiene una correlación positiva (directa) con Y_2.

Finalmente, se dan las coordenadas de cada variedad en las dos primeras componentes, así como una representación gráfica de las mismas.

	Y1	*Y2*
1	2.2441	-1.8570
2	3.5457	0.8433
3	4.2135	1.8231
4	4.1255	-2.8477
5	0.2910	2.6106
6	-2.0315	1.6109
7	-3.4699	-0.5631
8	-2.0417	-0.7475
9	-2.7221	0.5994
10	-4.1545	-1.3735



Atendiendo a esta representación gráfica, podemos decir que existen fundamentalmente dos grandes grupos: uno formado por las variedades 1, 2, 3 y 4 y otro formado por el resto de las variedades. El primer grupo caracterizado por los mejores rendimientos en las cuatro primeras condiciones de estrés (E1, E2, E3 y E4), en contraposición con el segundo grupo el cual presentó el peor comportamiento.

Ejemplo 2: En este caso se estudió el comportamiento de 18 variedades de boniato, a las que se le evaluaron un total de 10 variables:

Varie-dades	Peso (g)	Largo (cm)	Ancho (cm)	Pro-teína (%)	Sólidos Totales	Sólidos Soluble	Amino-ácidos (%)	Azúcares Totales	Ceniza (%)	Fibra (%)
1	191.2	11.0	4.5	1.99	47.16	9.2	37.69	5.95	1.46	1.28
2	150.4	13.5	5.0	1.75	37.14	9.0	28.14	4.39	1.53	1.51
3	155.7	15.0	6.5	1.91	35.65	9.5	29.35	4.43	1.25	1.26
4	175.8	16.0	4.5	1.89	33.73	8.8	25.22	3.62	0.94	1.42
5	160.0	14.5	3.5	2.03	31.81	10.0	25.45	4.12	1.31	1.02
6	146.7	15.0	4.0	2.59	38.76	10.4	27.65	5.09	1.09	1.73
7	231.3	20.5	5.5	1.94	33.76	10.0	26.22	3.92	0.79	1.19
8	172.5	12.5	5.0	1.21	31.17	10.8	23.39	3.65	0.75	1.36
9	240.2	24.0	7.0	2.14	32.73	12.0	24.63	4.56	0.82	1.01
10	151.7	15.0	3.5	1.90	30.91	8.2	25.64	2.83	1.07	1.21
11	220.1	20.5	3.5	1.64	31.25	8.3	24.61	3.02	0.70	1.19
12	167.2	15.0	6.5	1.29	31.55	8.7	24.63	3.29	1.31	1.31
13	202.2	20.5	4.0	1.85	33.38	11.2	27.62	4.86	1.13	1.10
14	204.1	17.0	4.5	1.64	36.11	10.6	28.73	4.83	1.21	1.38
15	217.8	17.5	4.5	1.22	31.11	9.4	19.90	3.38	1.01	1.25
16	241.7	16.0	3.5	2.19	40.44	10.3	31.02	4.55	0.97	1.66
17	237.8	23.5	6.0	1.65	31.41	10.2	24.19	3.98	0.83	1.11
18	176.8	15.5	4.0	1.73	30.21	10.0	23.61	3.71	0.82	1.04

Como puede observarse en este ejemplo, las variables que se evalúan están medidas en diferente escala, por lo que es necesario trabajar con la matriz de correlaciones, en lugar de hacerlo con la matriz de varianzas y covarianzas.

Resultados:

Matriz de correlaciones:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
X1	1.00									
X2	0.73	1.00								
X3	0.15	0.31	1.00							
X4	-0.06	0.05	-0.20	1.00						
X5	-0.02	-0.42	-0.11	0.50	1.00					
X6	0.39	0.39	0.25	0.24	0.02	1.00				
X7	-0.07	-0.40	-0.11	0.49	0.92	-0.02	1.00			
X8	0.05	-0.16	0.06	0.54	0.82	0.49	0.78	1.00		
X9	-0.54	-0.58	0.01	0.11	0.52	-0.25	0.56	0.46	1.00	
X10	-0.29	-0.45	-0.21	0.26	0.53	-0.14	0.33	0.25	0.24	1.00

Valores propios y varianza explicada:

	Valor propio	Varianza explicada	Varianza acumulada
1	3.9391	39.391	39.391
2	2.3869	23.869	63.261
3	1.1704	11.704	74.966

Vectores propios (coeficientes de la combinación lineal):

	Y1	*Y2*	*Y3*
X1	-0.097	-0.299	-0.153
X2	-0.163	-0.268	-0.057
X3	-0.056	-0.129	0.671
X4	0.133	-0.181	-0.332
X5	0.231	-0.110	-0.025
X6	-0.019	-0.327	0.086
X7	0.223	-0.100	0.038
X8	0.192	-0.241	0.139
X9	0.182	0.120	0.394
X10	0.148	0.068	-0.283

Correlaciones entre las variables originales y las componentes:

	Y1	*Y2*	*Y3*
X1	-0.385	**-0.715**	-0.180
X2	-0.642	-0.641	-0.067
X3	-0.221	-0.308	**0.786**
X4	0.527	-0.432	-0.389
X5	**0.913**	-0.264	-0.030
X6	-0.075	**-0.782**	0.100
X7	**0.881**	-0.240	0.045
X8	**0.758**	-0.575	0.163
X9	**0.718**	0.287	0.462
X10	0.584	0.163	-0.332

Valores de los individuos (18 variedades de boniato), en las tres primeras componentes:

Variedad	*Y1*	*Y2*	*Y3*
1	**2.438**	-0.420	0.717
2	**1.069**	0.982	0.854
3	0.580	0.183	**1.520**
4	-0.070	0.704	-0.702
5	0.136	0.546	0.003
6	**1.424**	-0.249	**-1.322**
7	-0.630	-0.917	-0.233
8	-0.647	0.639	0.269
9	-0.950	**-2.348**	0.940
10	-0.265	**1.589**	-0.856
11	**-1.094**	0.451	**-1.533**
12	-0.370	**1.283**	**1.794**
13	-0.054	**-1.006**	0.036
14	0.467	-0.476	0.260
15	**-1.113**	0.632	0.037
16	0.996	-0.919	**-1.858**
17	**-1.183**	**-1.081**	0.458
18	-0.733	0.407	-0.387

Como puede observarse, en este caso fue necesario trabajar con las tres primeras componentes, las cuales explican en su totalidad un 74.96 % de la información inicial.

La primera componente (la más importante) está caracterizada fundamentalmente por las variables *X5* (sólidos totales), *X7* (aminoácidos), *X8* (azúcares totales), y *X9* (cenizas).

La segunda componente está caracterizada por las variables *X1* (peso) y *X6* (sólidos solubles), mientras que la tercera componente está relacionada con el comportamiento de las variedades de boniato en la variable *X3* (ancho).

La primera componente contrapone las variedades 1, 2 y 6 de las variedades 11, 15 y 17. El primer grupo está caracterizado por presentar los mayores valores de sólidos totales, aminoácidos, azúcares totales y cenizas, mientras que el segundo grupo presenta los menores resultados en las variables antes mencionadas.

En el caso de la segunda componente, se contraponen las variedades 10 y 12 de las variedades 9, 13 y 17. El primer grupo está caracterizado por presentar los menores valores en el peso y los sólidos solubles (correlación negativa), mientras que el segundo grupo se distingue por valores de peso y sólidos solubles superiores al resto de las variedades.

Finalmente, la componente #3 contrapone las variedades 3 y 12 de las variedades 6, 11 y 16; el primer grupo caracterizado por presentar valores relativamente altos en cuanto al ancho, mientras que el segundo grupo tiene un comportamiento completamente opuesto.

Nótese que con la aplicación del Análisis de Componentes Principales, no solamente hemos caracterizado los individuos (variedades), sino que nos ha permitido además obtener información sobre relaciones de asociación entre variables. Varias variables se agrupan en una misma componente solamente, si están estrechamente correlacionadas.

♦ ***Otras aplicaciones en la agricultura***

Como se ha dicho anteriormente, el Análisis de Componentes Principales ha sido muy utilizado en el análisis de datos provenientes de investigaciones en la rama agrícola.

Le Minh Hong (1992), en su tesis de Doctorado, utilizó este método para seleccionar de un conjunto de cuatro muestreos correspondientes a diferentes períodos del desarrollo de la planta, el de mayor variabilidad, para con ello realizar posteriores análisis multivariados ya encaminados al muestreo seleccionado.

En el estudio se consideraron 15 variedades de tomate, a las que se les evaluaron un conjunto de seis caracteres de tipo fisiológico.

Los períodos de desarrollo de la planta analizados fueron: postura, floración, fructificación y maduración.

El autor concluyó con la aplicación del método que el período de mayor variabilidad era el correspondiente a la etapa de maduración, a lo cual da una explicación desde un punto de vista biológico en su trabajo de tesis.

En otro trabajo, Iglesias y Lourdes Iglesias (1995) utilizaron el Análisis de Componentes Principales, para clasificar 43 variedades de trigo atendiendo a su comportamiento en siete indicadores evaluados:

1- altura (A)
2- longitud de la espiga (LE)
3- granos llenos por espiga (G/E)
4- espigas por metro cuadrado (E/m^2)
5- peso de 1000 granos (P 1000)
6- rendimiento (R)
7- ciclo (C).

En este caso, se logró explicar con las tres primeras componentes un 76.7 % de la variabilidad total, y se clasificaron las variedades en tres grupos fundamentales, atendiendo al comportamiento de éstas en las componentes analizadas.

El método fue utilizado además por Noraida Pérez, Ismail y María C. González (1995), los que trabajaron con datos provenientes de 24 líneas de arroz, que se formaron a partir de cruces de las variedades Amistad-82 y 2077, sembradas en campo en parcelas de 6 m^2 y a las que se le evaluaron un conjunto de nueve caracteres. En este trabajo, el autor hizo una caracterización atendiendo al primer y tercer ejes o componentes principales, debido a que en ellos se agrupaban los caracteres más importantes dentro de su análisis.

Por otra parte, Plana (1991) utilizó el Análisis de Componentes Principales, con el objetivo de caracterizar el material de plantación proveniente de diferentes variedades, cepas y épocas de corte, en el cultivo de la caña de azúcar. Como índice de similitud para la interpretación de las componentes, se utilizó el coeficiente de correlación lineal estimado según Anderson (1968).

♦ ***Algunas consideraciones***

El Análisis de Componentes Principales resulta más efectivo en la medida en que inicialmente exista una estructura de correlación marcada entre las variables; si inicialmente hay poca correlación, el análisis deja de ser efectivo, ya que esto significaría que las componentes principales son prácticamente las variables iniciales; resulta muy difícil, en este caso, tratar de explicar con menos variables la información suministrada por las variables iniciales.

Por otra parte, en ocasiones se utiliza el Análisis de Componentes Principales, para formar grupos de individuos atendiendo a la representación gráfica de éstos en el plano formado por las dos primeras componentes. Al respecto, es oportuno señalar que si los dos primeros ejes no explican un por ciento considerablemente alto de la variación total, el agrupamiento formado a partir de la representación gráfica carecería de sentido. Además, en caso de que el por ciento de variabilidad extraído por los dos primeros ejes sea alto, se debe prestar atención al hecho de que en una representación gráfica donde los ejes coordenados son las componentes, las variaciones en dirección horizontal son más importantes que las variaciones cuando nos movemos en sentido vertical; por tanto, formar grupos de individuos a partir de su ubicación en el plano puede introducir errores. Se sugiere para estar seguros de la calidad del agrupamiento, combinar esta técnica con otros métodos multivariados, como el Análisis Factorial Discriminante, o simplemente utilizar un Cluster Análisis.

Entre los errores más frecuentes que se cometen en la aplicación del método, se puede citar el hecho de asociar una poca variabilidad explicada con las primeras componentes, con la poca variabilidad de las variables iniciales. En tal sentido, se puede decir que una situación no tiene ninguna relación con la otra; el no poder explicar con pocas componentes gran parte de la variabilidad o información inicial está más bien asociado con la poca correlación existente entre las variables iniciales.

Por último, es conveniente plantear que aprovechando el hecho de que el método construye variables incorrelacionadas, en ocasiones las componentes son usadas, con el objetivo de ser empleadas como variables independientes dentro de un modelo de Regresión Lineal Múltiple. Esto responde al hecho de protegerse sobre la presencia de multicolinealidad (Cuthbert, 1980), la cual aumenta las varianzas de las estimaciones de los parámetros en un modelo y se presenta cuando exísten estructuras de correlación muy marcadas en las variables independientes.

Análisis factorial discriminante

♦ ***Introducción***

El Análisis Factorial Discriminante es un método multivariado, que permite describir diferencias entre grupos o efectos de tratamientos, a partir de un valor heurístico calculado de las mejores funciones lineales del vector variable o ejes discriminantes (CooLey, 1971).

En este caso, nuevamente se parte de una matriz X_{nxp} de individuos-variables, aunque con la diferencia fundamental de que exíste una partición por filas de la matriz de datos, es decir, en la muestra de individuos existen grupos formados *a priori*.

El objetivo del análisis es tratar de describir a partir de los ejes discriminantes, las diferencias que puedan existir entre estos grupos, así como reflejar las variables que mayor aporte hacen a esta diferenciación. Interrogantes tales como saber si el conjunto de variables incluidas en el análisis contribuyen a la diferenciación de los grupos, pueden ser contestadas a partir de la aplicación de este método.

Es frecuente ver la aplicación del Análisis Factorial Discriminante como complemento de otros métodos multivariados, tales como Análisis de Componentes Principales y Cluster Análisis, constituyendo una de las técnicas multivariadas de mayor aplicación.

♦ ***Nociones del fundamento matemático***

El fundamento matemático del método consiste en hacer uso del teorema de Fischer para la descomposición de la suma de cuadrados del ANOVA en componentes aditivas (Cooley, 1971).

En este caso, como en todo análisis multivariado, se parte de una matriz de datos *X* de orden *nxp*:

$$X=\begin{bmatrix} & \vdots & \\ \cdots & ij & \cdots \\ & \vdots & \end{bmatrix}_{nxp}$$

n: # de individuos
p: # de variables
n≥p

Lo distintivo del método es que la matriz *X* aparece particionada por filas, es decir, existen grupos formados *a priori*, así el elemento x_{ijk} representa el valor que toma en la variable *j* el individuo *i* del grupo *k*.

Sea *g* el número de grupos y n_i la cantidad de individuos de la muestra que pertenecen al grupo *i*. Entonces:

$$n = \sum_{i=1}^{g} n_i \; 0$$

A partir de la matriz inicial *X* se definen las matrices *T*, *A* y *W*, todas ellas de orden *p*, las cuales van a representar las matrices de varianza y covarianza total, matriz de varianza y covarianza entre grupos, y matriz de varianza y covarianza dentro de grupos, respectivamente.

$$T=(t_{ls}) \qquad l=1..j\ ,\ s=1..j$$

$A=(a_{ls})$

$W=(w_{ls})$

Siendo: $t_{ls} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ilk} - m_l)(x_{isk} - m_s)' / (n - 1)$

$$a_{ls} = g \sum_{k=1}^{g} (mg_{lk} - m_l)(mg_{sk} - m_s)' / (g-1)$$

$$w_{ls} = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (x_{ilk} - mg_{lk})(x_{isk} - mg_{sk})' / (n - g)$$

donde m_j es la media general de los datos en la variable *j*, y mg_{jk} representa la media para la variable *j* en el grupo *k*.

Se trata entonces de encontrar *g-1* funciones discriminantes ($Y_1, Y_2, ..., Y_{g-1}$), que se forman como combinaciones lineales de las variables originales y que deben tener la característica de discriminar lo máximo posible los *g* grupos existentes, además de estar incorrelacionadas.

Sean:

$$Y_1 = \sum_{j=1}^{p} \alpha_{1j} X_j$$

$$Y_2 = \sum_{j=1}^{p} \alpha_{2j} X_j$$

.
.
.

$$Y_{g-1} = \sum_{j=1}^{p} \alpha_{g-1\,j} X_j$$

El problema de encontrar las funciones discriminantes con las bondades antes mencionadas, se traduce en buscar los valores y vectores propios de la matriz $T^{-1} W$.

Sean $l_1, l_2, ..., l_{g-1}$, los *g-1* valores propios de $T^{-1} W$, entonces los coeficientes α_{1j} *j=1..p*, que definen al primer eje discriminante serán las coordenadas del primer vector propio, es decir, del vector propio asociado al mayor valor propio (l_1), y así sucesivamente, o sea, los coeficientes α_{g-1j} *j=1..p*, son las coordenadas del vector propio asociado al valor propio l_{g-1}.

Cada eje extrae una parte de la inercia total o información inicial, así, por ejemplo, la inercia asociada a Y_1 será:

$$(l_1 / \sum_{k=1}^{g-1} l_k) * 100\%.$$

En este método, al igual que en el Análisis de Componentes Principales, no resulta necesario explicar los *g-1* ejes discriminantes o variables canónicas, puede incluso ocurrir que ningún eje discriminante sea importante en la diferenciación de los grupos; en tal caso, concluimos diciendo que no existen diferencias entre los grupos atendiendo al conjunto de variables analizadas.

Puede presentarse el caso de que una sola variable canónica sea suficiente, para explicar las diferencias entre los grupos, incluso pudiera ocurrir que sea necesario considerar las *g-1* funciones discriminantes.

¿Cómo saber cuántos ejes discriminantes o variables canónicas debemos considerar?

Primeramente se analiza si es necesaria la inclusión de al menos una variable canónica o eje discriminante; si la respuesta es afirmativa, se pasa a considerar si es necesaria la inclusión de una segunda variable canónica. Este algorítmo se realiza hasta obtener una respuesta negativa; en tal caso, se considera solamente la cantidad de ejes incluidos hasta ese momento.

Si la respuesta a la primera interrogante resultara negativa, indica que ningún eje discriminante es importante en la diferenciación de los grupos, o lo que es lo mismo, no existen diferencias entre los grupos.0

Para dar respuesta a cada una de estas interrogantes, se realizan pruebas X^2 sucesivas; de la siguiente manera :

Para saber si es importante o no considerar la primera variable canónica, se calculan las matrices W_0 y T_0, correspondientes a las matrices de varianzas y covarianzas dentro de grupos y totales, para las *g-1* funciones discriminantes o variables canónicas.

Posteriormente, se calcula el λ_0 de Wilks, como el determinante del cociente de W_0/T_0:

$$\lambda_0 = DET(W_0/T_0)$$

A partir del valor de λ_0 , se calcula la $X^2 = -\{N-1-\frac{1}{2}(p+1)*log(\lambda_0)\}$ según Judez (1989).

Este valor se compara con una X^2 con $p(g-1)$ grados de libertad. Si el valor calculado es mayor que el tabulado, se concluye que la primera variable canónica resulta importante en la diferenciación de los grupos; en caso contrario, no se considera este primer eje discriminante y, por tanto, no se considera ninguno de los restantes, por ser menos importante que el primero en la diferenciación de los grupos.

$$(\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{g-1})$$

En caso de incluir el primer eje, como se dijo anteriormente, se pasa a realizar la prueba de significación para el segundo eje discriminante; en este caso, se calculan las matrices W_1 y T_1 correspondientes a las matrices de varianzas y covarianzas dentro de grupos y totales respectivamente; pero en este caso, asociadas a las *g-2* funciones discriminantes restantes (se elimina la primera).

Nuevamente se calcula el Lambda de Wilks, λ_1 como el $det(W_1/T_1)$ y la X^2 asociada. En este caso, el valor calculado se compara con una X^2 con $(p-1)*(g-2)$ grados de libertad.

Se repite el procedimiento hasta la etapa en que se rechace la inclusión de un eje discriminante; en tal caso, se trabaja solamente con el número de ejes considerados hasta esa etapa.

Una vez determinado el número de funciones o ejes discriminantes que son necesarios explicar, se pasa, al igual que en las Componentes Principales, a determinar dentro de cada eje cuáles de las variables originales son las que caracterizan la función discriminante. Esto se determina mediante el coeficiente de correlación dentro de grupos entre cada eje discriminante con las variables originales.

El análisis, además, aporta un criterio utilizado para saber si las variables iniciales discriminan los grupos existentes; el mismo consiste en evaluar los *p* valores iniciales de cada individuo en la primera función discriminante Y_1, el valor obtenido sirve para reclasificar cada individuo en un nuevo grupo, atendiendo a su proximidad respecto a los centros de gravedad de cada grupo. En cada caso un individuo puede ser reubicado en el mismo grupo de pertenencia o, por el contrario, puede ser ubicado en otro grupo al cual no pertenece. Si las variables analizadas discriminan de forma eficiente los grupos, el por ciento de individuos mal clasificados será mínimo, indicando que existen diferencias entre los grupos analizados.

♦ ***Ejemplo de aplicación***

En un experimento de campo, se probaron los efectos de tres dosis de fertilizante sobre cuatro indicadores del rendimiento en el cultivo del arroz (I1,I2,I3 y I4). Se utilizó un diseño Completamente Aleatorizado desbalanceado, por lo que se tomaron tres, tres y cuatro observaciones por tratamiento respectivamente. Cada tratamiento con sus observaciones son considerados como un grupo, debido a que se quiere conocer si existen diferencias entre los tratamientos, atendiendo a su comportamiento en los cuatro indicadores evaluados, a partir del uso del Análisis Factorial Discriminante.

Matriz inicial de datos:

	I1	**I2**	**I3**	**I4**
1	45.00	1.56	13.00	14.00
2	50.00	1.60	12.00	16.00
3	50.00	1.65	13.00	15.00
4	60.00	1.75	15.00	9.00
5	60.00	1.70	14.00	10.00
6	65.00	1.70	14.00	7.00
7	70.00	1.60	15.00	8.00
8	65.00	1.60	13.00	13.00
9	60.00	1.55	15.00	17.00
10	65.00	1.70	14.00	11.00

Esta matriz está particionada por filas:
Del individuo 1 al 3 Grupo 1
Del individuo 4 al 6 Grupo 2
Del individuo 7 al 10 Grupo 3

Por tanto, en este caso n=10, n_1=3, n_2=3, n_3=4, g=3 y p=4

Valores medio por grupos y caracteres:

Grupo 1	I1	48.33	Grupo 2	I1	61.66
	I2	1.60		I2	1.71
	I3	12.66		I3	14.33
	I4	15.00		I4	8.66
Grupo 3	I1	65.00			
	I2	1.61			
	I3	14.25			
	I4	12.25			

Matriz de varianzas y covarianzas entre grupos (A):

	I1	I2	I3	I4
I1	485.83	1.28	32.00	-100.00
I2		0.02	0.21	-1.13
I3			2.93	-12.33
I4				60.58

Matriz de varianzas y covarianzas dentro de grupos (W):

	I1	I2	I3	I4
I1	66.66	0.66	-5.00	-40.00
I2		0.02	-0.05	-0.26
I3			2.66	4.33
I4				49.41

Valores y vectores propios de $T^{-1}\,W$:

Eje	Valor propio	Contribución
1	14.46	87.1 %
2	2.13	12.9 %

Vectores propios:

Variables	Eje 1	Eje 2
I1	-1.04	0.493
I2	0.019	-0.636
I3	-0.241	-0.266
I4	-0.395	0.541

A partir de los vectores propios se construyen los ejes discriminantes *Y1* y *Y2*:
Y1= -1.0289I1+0.1131I2-0.1726I3-0.2143I4
Y2= 0.5207I1-0.4523I2-0.4675I3+0.6004I4

Valor propio	Inercia	Chi-cuadrado	Grados de libertad	Significación (%)
14.46	87.1 %	21.34	8	0.64
3.132	12.9 %	6.28	3	9.71

A partir del análisis de los valores propios, se concluye que ambos ejes aportan en la diferenciación de los grupos; es importante explicar cada uno de ellos.

Podemos concluir, además, diciendo que como al menos un valor propio fue significativo (en este caso los dos), existen diferencias entre los tres tratamientos estudiados, a partir de su comportamiento en los indicadores analizados.

Correlaciones entre grupos entre las variables originales y los ejes discriminantes:

Variables	Eje1	Eje2
I1	**-0.9807**	-0.1956
I2	-0.1271	**-0.9919**
I3	**-0.9054**	-0.4245
I4	0.5071	**0.8619**

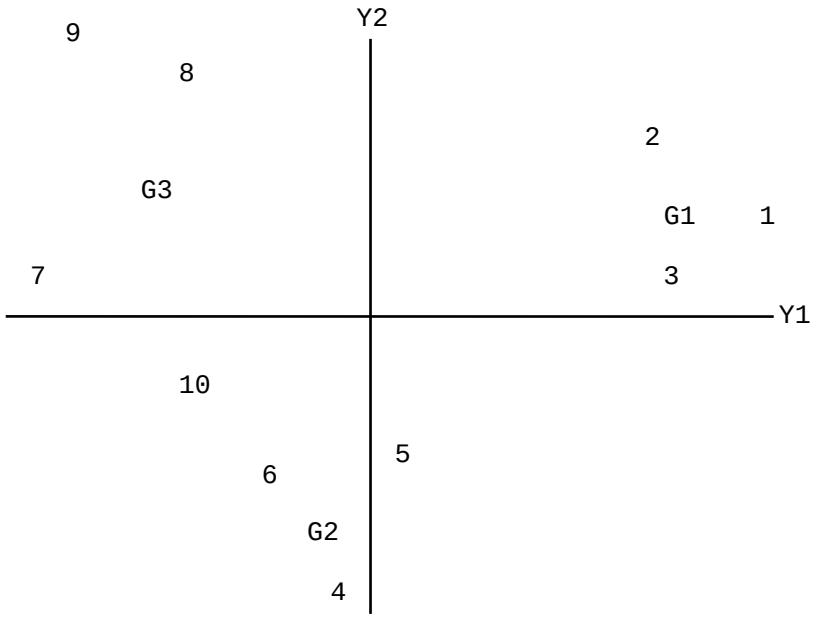
A partir del análisis de esta tabla de correlaciones, se concluye que el Eje 1 diferencia los tratamientos atendiendo a su comportamiento en los indicadores I1 y I3; siendo esta la diferencia más marcada que se establece entre los grupos. Por otra parte, el eje 2 discrimina a los grupos atendiendo a su comportamiento en los indicadores I2 e I4.

Tabla de pertenencia:

	Grupo de afectación		
Grupo de pertenencia	1	2	3
1	3	0	0
2	0	3	0
3	0	0	4

Al analizar esta tabla, podemos decir que el 100 % de los individuos se clasificaron correctamente a partir del primer eje discriminante, siendo esto una evidencia más acerca del efecto diferenciado que provocan las distintas dosis de fertilizante.

Representación gráfica de los individuos y centros de gravedad de los grupos:



El grupo 1 (dosis 1) al estar más a la derecha, es el que presenta los menores valores de los indicadores I1 e I3, debido a que la correlación de las variables I1 e I3 con *Y1* es negativa.

Por otra parte, se puede decir que el grupo 2 (dosis 2) se caracteriza por presentar valores intermedios de los indicadores 1 y 3, mientras que el grupo 3 (dosis 3) presenta los valores más altos de estos indicadores.

Refiriéndonos a la segunda variable canónica, podemos decir que esta diferencia los grupos atendiendo a su comportamiento en los indicadores 2 y 4 fundamentalmente (ver correlaciones); de esta forma, podemos decir que los tratamientos 1 y 3 presentan los mayores valores del indicador 2 y los menores valores del indicador 4, en contraposición al tratamiento 2.

Finalmente, como resultado del análisis se concluye que los tres tratamientos estudiados, es decir, las tres dosis de fertilizante, tienen un efecto diferenciado en los indicadores 1 y 3 fundamentalmente, siendo la dosis 3 la de mejores resultados.

♦ ***Otras aplicaciones en la agricultura***

En un estudio realizado en el cultivo del tomate, con el propósito de disminuir los niveles de riego en este cultivo, Dell'Amico (1992) utilizó el Análisis Factorial Discriminante. En el trabajo se analizaron cuatro tratamientos, que consistieron en regar con diferentes dosis de agua.

Se midió el efecto de estos tratamientos sobre seis variables fundamentales: rendimiento, pH, contenido relativo de agua, prolina, actividad nitrato reductasa y clorofila.

Como resultado del análisis, el autor demostró que independientemente que los tratamientos ejercieron un efecto diferenciado en las variables analizadas, el tratamiento control, es decir, el que menos agua suministraba, no difirió del tratamiento correspondiente a las normas técnicas, permitiendo esto recomendar una tecnología de riego con menos agua.

En otro trabajo, Cruz (1995) utilizó el Análisis Factorial Discriminante, para corroborar acerca de un agrupamiento realizado a partir de un Análisis de Componentes Principales; el autor en un estudio previo pudo hacer un agrupamiento de sus individuos (en este caso variedades de caña), el cual estaba en correspondencia con el progenitor madre de cada variedad, es decir, todos los individuos cuya madre era la misma se agruparon, sin importar la procedencia respecto al padre.

Al realizar el correspondiente Análisis Factorial Discriminante, el autor obtuvo un 95 % de buena clasificación, corroborando por tanto que sus grupos estaban bien constituidos, por lo que se demostró en el trabajo el gran efecto materno exístente.

Los caracteres analizados en este caso fueron vigor, brix, diámetro y longitud del tallo.

- ***Algunas consideraciones***

La aplicación del Análisis Factorial Discrimiminante a diferencia del Análisis de Componentes Principales siempre es efectiva, ya que el mismo va a facilitar información acerca de si las variables analizadas discriminan los grupos analizados, o lo que es lo mismo, si los grupos considerados tienen efectos diferenciados en el conjunto de variables medidas. Su uso correcto siempre va a aportar alguna información del problema analizado.

Es importante siempre que se considere más de un eje para la interpretación de los resultados, tener en cuenta que el primero de ellos es el más importante, puesto que está asociado al mayor valor propio. Este resultado es de vital importancia, ya que por ejemplo en una representación gráfica se olvida que diferencias en sentido horizontal (*Y1*) son más importantes que diferencias en sentido vertical (*Y2*) y, en estos casos, se arriba a interpretaciones alejadas de lo verdadero.

En el caso del Análisis Factorial Discriminante, al igual que en el Análisis de Varianza clásico, puede darse el caso que no se detecten diferencias entre grupos, debido a la gran variabilidad existente dentro de ellos, independientemente que los valores medio por grupo sean muy disímiles. En estos casos, no debemos concluir diciendo que no existen diferencias entre grupos; en realidad, lo más conveniente sería volver a conformar los grupos creados, ya que ellos no responden al concepto de grupo.

Esta técnica de Análisis es frecuente verla utilizada, como complemento de otros métodos multivariados empleados, con el propósito de la formación de grupos, como por ejemplo Análisis de Componentes Principales y Cluster Análisis. Es a partir de ella que puede ser verificado lo correcto o no de determinado agrupamiento.

Análisis discriminante

♦ *Introducción*

El investigador se encuentra a menudo con el problema de clasificar los elementos o grupos de un universo, según un criterio definido con más de una característica. No es raro ver en problemas de este tipo, la aplicación de una técnica primitiva de clasificación, usando ponderaciones distintas para cada una de las características, según la importancia que tengan a juicio o criterio del investigador. De esta forma, las ponderaciones dadas a las características son subjetivas y no pueden considerarse satisfactorias y adecuadas (Marta Flores, 1972).

El Análisis Discriminante es un método para hacer una clasificación objetiva de unidades o grupos de unidades, según el criterio definido por varias características. Esta clasificación o rango se establece por medio de un número (sintético) integrado por las propiedades de todas las características consideradas.

Las primeras ideas de esta técnica multivariada, surgen en la cuarta década del siglo XX, relacionadas con algunas investigaciones biológicas y antropométricas. Mahalanobis (1930) y Fischer (1936) son los fundadores principales del Análisis Discriminante.

Gran importancia tiene el método en la solución de problemas, en los que queremos pronosticar acerca del comportamiento de determinadas variedades en una característica o variable cualitativa. Este pronóstico se hace a partir de la información suministrada por una serie de evaluaciones continuas que se realizan, las cuales el investigador incluye en el análisis, debido a que sospecha que puedan tener relación con el fenómeno estudiado.

El Análisis Discriminante puede utilizarse para separar universos, para reconocer si un elemento pertenece a universos dados y para clasificar universos según una o más características.

♦ *Nociones del fundamento matemático*

Como en todo análisis multivariado se parte de una matriz X_{nxp}, donde n representa el número de individuos de la muestra y p la cantidad de variables contínuas en ellos observadas.

En este caso, al igual que en el Análisis Factorial Discriminante, la matriz X va a estar particionada por filas; es decir, existen grupos de individuos formados *a priori*.

La partición va a estar dada a través de una variable de clasificación o efecto dependiente de las variables continuas observadas.

El objetivo del análisis es estudiar si es posible predecir a partir del comportamiento de los individuos en las p variables continuas en ellos observadas, cuál va a ser la categoría de la variable dependiente o de clasificación a la que él va a pertenecer.

Para ello se construyen g funciones discriminantes, una para cada grupo:

Sean: $Y1= \beta_1+\alpha_{11}X_1+\alpha_{12}X_2+.....+\alpha_{1p}X_p$

$Y2= \beta_2+\alpha_{21}X_1+\alpha_{22}X_2+.....\alpha_{2p}X_p$

. .

. .

. .

$Yg= \beta_g+\alpha_{g1}X_1+\alpha_{g2}X_2+.....+\alpha_{gp}X_p$

El problema de encontrar dichas funciones, se traduce en encontrar los coeficientes β_i *y* α_{ij}, *i=1..g j=1..p.*

Estos coeficientes deben buscarse de tal forma, que se haga mínima la varianza dentro de grupos y máxima la varianza entre grupos.

Se demuestra que los coeficientes con tales características son :

$\beta_i= -½\ \mu_i' S^{-1}\mu_i \qquad i=1..g$

$$\begin{bmatrix} \alpha i1 \\ . \\ . \\ \alpha ip \end{bmatrix} = \mu_i' S^{-1}$$

donde (μ_i) es el vector media para el grupo *i* y *S* es la matriz de varianzas y covarianzas ponderada:

$S= (\Sigma_i (n_i-1)S_i) / \Sigma_i (n_i-g)$, siendo S_i la matriz de varianzas y covarianzas para el grupo *i.*

Las *g* funciones discriminantes tendrán un buen poder de diferenciación entre los grupos, en la medida que inicialmente existan diferencias entre ellos.

Precisamente, para conocer si existen o no diferencias entre los grupos, atendiendo a su comportamiento en el conjunto de las *p* variables continuas en ellos observadas, se calcula el Lambda de Wilks, definido como:

$\lambda= det(W)/det(T)$, siendo *W* y *T* las matrices de varianzas y covarianzas dentro de grupos y total, respectivamente.

Este valor se encuentra entre 0 y 1: un valor de cero indica discriminación perfecta, mientras que un valor de uno significa ausencia de diferencias entre los grupos. Para conocer si existen diferencias entre los grupos, el Lambda de Wilks calculado se aproxima a una *F* de Fisher, según Rao (1952).

En caso de que existan diferencias entre los grupos, resulta interesante conocer qué variables son responsables de esta diferenciación. Para ello, se calcula para cada carácter el Lambda parcial, con su respectiva significación.

λ parcial= (λ después)/(λ antes), donde el valor λ después significa el valor de λ que se calcula con todas las variables, mientras que λ antes es el que se calcula sin incluir la variable que se está analizando. En realidad, el valor de λ parcial significa cuán importante resulta la variable en la diferenciación de los grupos; en este caso, si el λ parcial se aproxima a uno, significa que introducir la variable no aportaría nada en la diferenciación de los grupos; mientras que un valor de λ parcial próximo a cero, significa que al introducir la variable, la diferenciación entre los grupos se hace más acentuada.

Para realizar la significación para el λ parcial, se calcula para cada variable el estadígrafo:

$$F=[(n-g-p+1)/(g-1)]*[(1-\lambda\ parcial)/(\lambda\ parcial)]$$

el cual se distribuye *F* de Fisher con *g-1* y *n-g-p+1* grados de libertad en el numerador y denominador respectivamente.

Otro criterio muy utilizado para saber si existen diferencias entre los grupos, atendiendo a su comportamiento en el conjunto de las *p* variables continuas en ellos observadas, es el llamado por ciento de buena clasificación, que es el resultado de reclasificar cada uno de los individuos de la muestra en un nuevo grupo, haciendo uso de las *g* funciones discriminantes.

Para ello, se calcula para cada individuo de la muestra su valor en cada una de las *g* funciones discriminantes, el individuo se clasificará en el grupo para el cual el valor de la función discriminante asociada haya sido mayor, finalmente se obtiene el por ciento de individuos bien clasificados como resultado de esta reclasificación. En la medida que este por ciento se aproxime a 100, será un indicador de que existen diferencias entre los grupos y, por tanto, indica que las funciones discriminantes pueden ser usadas para clasificar un individuo ajeno a la muestra, en uno de los grupos estudiados.

Es muy importante saber, que el procedimiento para clasificar un individuo ajeno a la muestra en uno de los *g* grupos, será efectivo solamente si existen diferencias entre los grupos, en caso contrario ello carecería de sentido.

♦ ***Ejemplo de aplicación***

En este caso, tomaremos un juego de datos en el que se evaluaron en dos variedades de yuca (Señorita y CMC-40) un conjunto de ocho variables:

X1- altura (cm)
X2- largo (cm)
X3- diámetro (cm)
X4- número
X5- masa por planta (kg)
X6- biomasa por planta (kg)
X7- biomasa por hectárea ($t.ha^{-1}$)
X8- rendimiento ($t.ha^{-1}$)

El experimento se realizó durante tres años, y en cada año se recogen nueve observaciones por variedad; esto da como resultado un total de 27 observaciones o individuos por variedad, para un tamaño de muestra total de 54.

Tratamiento	Altura	Largo	Diámetro	Número	Masa por planta	Biomasa por planta	Biomasa. ha^{-1}	Rendimiento	Variedad
1	1.66	30.14	4.62	4.88	2.09	2.13	23.41	21.00	1
2	1.64	30.72	4.56	4.29	1.86	1.74	16.01	18.08	1
3	1.63	30.89	5.12	4.94	2.42	2.16	22.80	32.65	1
4	1.62	29.30	5.08	5.20	1.54	1.47	15.54	28.62	1
5	1.61	31.65	5.34	4.62	3.77	2.14	23.37	33.00	1
6	1.60	30.78	4.72	4.48	1.68	1.79	20.28	31.02	1
7	1.57	29.14	5.34	4.80	2.55	1.44	15.53	28.74	1
8	1.57	29.82	5.60	4.20	2.70	3.33	34.26	38.71	1
9	1.51	29.74	5.56	4.23	3.27	3.53	39.02	38.48	1
10	1.51	29.54	4.66	5.57	1.70	1.26	12.88	28.83	1
11	1.65	27.88	4.27	4.65	1.49	1.32	14.25	24.80	1
12	1.60	29.08	4.52	5.50	2.06	1.29	13.94	32.14	1
13	1.70	28.84	4.36	5.30	2.01	1.48	14.03	26.50	1
14	1.69	29.85	4.35	5.35	2.35	1.44	17.18	33.31	1
15	1.64	29.13	4.44	5.45	2.04	1.55	18.46	28.28	1
16	1.58	28.26	4.59	5.40	2.26	1.35	14.74	26.84	1
17	1.57	25.63	5.06	5.60	2.44	1.81	21.10	32.12	1
18	1.61	29.10	4.79	5.50	2.66	1.62	15.56	33.03	1
19	1.76	29.42	6.15	4.86	2.94	2.18	23.00	33.48	1
20	1.78	29.76	6.29	3.99	2.68	1.67	17.30	28.76	1
21	1.84	33.17	6.40	4.19	4.57	2.82	29.26	33.96	1
22	1.74	38.20	5.83	4.94	2.69	1.58	19.97	28.90	1
23	1.63	35.73	5.54	4.29	2.42	1.21	13.26	34.43	1
24	1.64	31.83	5.60	4.92	1.86	1.42	12.71	31.02	1
25	1.75	33.51	6.60	5.15	3.72	2.77	29.21	30.86	1
26	1.90	32.33	6.42	6.17	3.25	2.27	23.94	34.05	1
27	1.53	34.76	5.35	5.50	2.53	2.90	19.64	33.84	1
28	1.60	27.62	5.02	4.88	2.47	0.73	7.42	24.83	2
29	1.61	27.69	4.83	4.29	1.68	0.79	8.17	18.82	2
30	1.62	26.72	4.72	4.94	2.17	0.72	12.41	27.43	2
31	1.57	26.26	4.75	5.20	1.67	1.15	11.97	19.16	2
32	1.63	26.06	5.17	4.62	2.50	1.38	11.02	28.81	2
33	1.64	24.30	4.65	4.48	1.80	0.97	8.92	21.24	2
34	1.66	25.30	4.61	4.80	1.76	1.17	12.31	20.46	2
35	1.57	29.34	5.78	4.20	2.41	1.75	19.35	28.36	2

36	1.51	25.48	5.35	4.23	2.29	1.77	19.99	25.36	2
37	1.32	27.91	4.24	6.15	1.55	1.55	18.45	29.10	2
38	1.38	31.66	5.63	4.94	1.73	1.73	16.85	24.97	2
39	1.32	28.98	4.86	4.95	1.60	1.60	17.45	28.66	2
40	1.39	31.99	4.71	5.31	1.76	1.76	16.94	28.67	2
41	1.35	28.29	4.47	4.55	1.45	1.45	14.94	28.55	2
42	1.30	22.58	3.59	4.95	1.36	1.36	17.32	26.21	2
43	1.36	28.90	4.74	5.50	1.68	1.68	16.81	26.76	2
44	1.37	30.15	5.33	4.90	1.63	1.63	17.55	26.10	2
45	1.36	30.40	5.26	5.62	1.55	1.55	16.88	28.63	2
46	1.46	27.73	4.62	5.65	2.34	1.15	12.42	24.72	2
47	1.49	29.50	5.05	5.05	1.96	1.27	13.37	21.02	2
48	1.46	34.73	4.32	5.10	2.33	1.37	14.41	25.15	2
49	1.48	35.84	4.27	5.35	1.47	1.56	15.81	15.43	2
50	1.49	37.76	4.32	4.12	1.37	1.21	13.72	24.85	2
51	1.47	32.33	4.35	4.97	1.73	1.32	11.64	77.72	2
52	1.32	36.66	4.47	5.27	1.87	1.60	16.77	20.09	2
53	1.47	33.78	4.39	5.20	1.65	1.62	16.77	27.93	2
54	1.44	36.22	4.79	5.60	1.72	1.62	17.44	28.19	2

La variable *X9* (Variedad) se refiere a la variedad a la que pertenece cada observación:

X9=1 indica que pertenece a la variedad CMC-40 (Grupo I)

X9=2 indica que pertenece a la variedad señorita (Grupo II)

Se quiere determinar si existe o no un efecto del factor variedad en el conjunto de las ocho variables observadas.

Resultados:

Matriz de varianza y covarianza dentro de grupos (W)

	X1	X2	X3	X4	X5	X6	X7	X8
X1	0.01	-0.01	0.02	-0.01	0.03	-0.01	-0.01	-0.01
X2	-0.03	11.50	0.38	0.07	0.20	0.40	2.90	0.20
X3	0.02	0.40	0.37	-0.08	0.23	0.14	1.30	1.00
X4	-0.01	0.10	-0.08	0.27	-0.06	-0.03	-0.40	0.10
X5	0.03	0.20	0.23	-0.06	0.33	0.14	1.50	0.90
X6	-0.01	0.40	0.14	-0.03	0.14	0.26	2.50	1.00
X7	-0.10	2.90	1.31	-0.39	1.48	2.50	28.20	11.00
X8	-0.10	0.20	0.98	0.08	0.95	0.99	11.00	18.50

Matriz de varianza y covarianza total (T)

	X1	X2	X3	X4	X5	X6	X7	X8
X1	0.02	0.00	0.05	-0.02	0.06	0.01	0.10	0.20
X2	0.01	11.50	0.48	0.06	0.35	0.51	1.10	1.50
X3	0.05	0.50	0.42	-0.08	0.31	0.20	1.90	1.70
X4	-0.02	0.10	-0.08	0.27	-0.06	-0.04	-0.40	0.00
X5	0.06	0.40	0.31	-0.06	0.43	0.23	2.40	0.90
X6	0.01	0.50	0.20	-0.04	0.23	0.33	3.20	1.80
X7	0.15	4.10	1.93	-0.42	2.36	3.17	34.90	18.80
X8	0.18	1.50	1.68	0.04	1.93	1.76	18.80	27.00

A partir de las matrices *T* y *W* se calcula el λ de Wilks, para conocer si existen diferencias entre las variedades atendiendo a su comportamiento en los ocho caracteres evaluados.

λ= determinante(*w*)/determinante(*T*); en nuestro caso:

λ= *0.2849*. Este valor se aproxima a una *F(8,45)=14.1159*, la cual es altamente significativa al ser comparada con la *F* de Fisher tabulada correspondiente.

Podemos decir, por tanto, que existe un efecto diferenciado entre las dos variedades de yuca, atendiendo a su comportamiento en los indicadores evaluados.

¿En qué variables se dan tales diferencias entre las dos variedades?.

Para responder la pregunta anterior, se hace necesario calcular para cada variable el λ-parcial.

	λ antes (Sin Xi)	λ-parcial	F(1,45)	Significación
Variable				
X1	0.6074	0.4690	50.930	***
X2	0.2903	0.9815	0.847	n.s
X3	0.3161	0.9011	4.936	**
X4	0.2884	0.9879	0.549	n.s
X5	0.2972	0.9586	1.942	n.s
X6	0.2966	0.9606	1.845	n.s
X7	0.2849	0.9999	0.003	n.s
X8	0.3844	0.7411	15.710	***

La primera columna de la tabla anterior se refiere al λ de Wilks, calculado sin incluir la variable en cuestión; no es más que el valor de (λ-antes) que aparece en la fórmula para el cálculo del λ-parcial por variable.

En la medida que este valor sea superior al λ de Wilks, calculado para todas las variables, será un indicio de lo importante que resulta el carácter analizado en la diferenciación de los grupos.

En nuestro caso, podemos ver cómo los caracteres *X1, X8* y *X3*, es decir, altura, rendimiento y diámetro, son las variables en las que existe un efecto diferenciado de las dos variedades de yuca que se estudian.

Valores medio por grupos y caracteres

	X1	X2	X3	X4	X5	X6	X7	X8
Grupo I	1.65	30.67	5.22	4.96	2.54	1.91	20.02	30.60
Grupo II	1.46	29.78	4.75	4.99	1.83	1.38	14.70	24.71

Como quiera que ya conocemos que las variables *X1, X3* y *X8*, son las más importantes en la diferenciación de las variedades, podemos discutir la tabla anterior, haciendo referencia solamente a estos caracteres.

Podemos plantear que la variedad CMC-40 es la que presenta la mayor altura, el mayor rendimiento y el mayor diámetro, en comparación con la variedad de yuca Señorita.

Por ciento de buena clasificación:

	Bien clasificados	Mal clasificados
Grupo I	25	2
Grupo II	2	25

% de buena clasificación: 92.59 %

Individuos mal clasificados: 7, 10, 32 y 53

Los dos primeros pertenecen a la variedad CMC-40 y, sin embargo, presentan las características de la otra variedad; por otra parte, los individuos 32 y 53, que corresponden ambos a la variedad Señorita, tienen un comportamiento propio de los individuos de la variedad CMC-40.

Otro enfoque del Análisis Discriminante. Ordenamiento de una muestra multivariada. Muchas son las ocasiones en que los análisis univariados independientes para cada variable, aportan que los individuos de mejor comportamiento en una característica, no son los de mejor comportamiento en la otra variable. En estos casos, no se sabe cómo proceder para seleccionar el de mejores resultados globales.

El objetivo del análisis es establecer un ordenamiento de los individuos, atendiendo a su comportamiento en el conjunto de las p variables observadas.

Para lograr el objetivo, se construye un individuo ficticio, cuyos valores en las p variables son los peores valores tomados en la muestra para cada una de las características evaluadas.

Sea $(z_{p1},z_{p2},...,z_{pp})$ las coordenadas de este individuo ficticio; así z_{p1} corresponderá al peor valor de la variable X_1 que aparece en la muestra de n individuos; z_{p2} representa el peor valor de la variable X_2 que se alcanza en la muestra y, así sucesivamente; z_{pp} representa el peor valor para la variable X_p. Nótese que estos valores no tienen por qué darse en el mismo individuo de la muestra.

El método consiste en calcular la distancia de cada uno de los individuos de la muestra a este individuo ficticio, formado con las peores características. Para calcular este valor, se debe hacer uso de algún índice de similitud o distancia.

Entre las distancias estadísticas más utilizadas se encuentran la Euclideana, la de Mahalanobis, la de Frechet, la Absoluta, la I-Distancia y la A-Distancia, entre otras. La elección de una de ellas debe estar en correspondencia con la naturaleza de los datos, estructuras de correlación y escalas de medición de variables.

Una vez seleccionada la distancia estadística a utilizar, se transforma la matriz X en un vector de orden n, en el cual va a aparecer en la posición i, la distancia del individuo i de la muestra al individuo ficticio. De esta forma, aquel individuo cuyo valor de distancia sea mayor, será el de mejor comportamiento, debido a que dista más de las peores características.

Ejemplo de aplicación: Se tomaron un conjunto de 18 variedades, a las cuales se les aplicaron tres estimuladores del crecimiento diferentes. Se ofrecen los datos de incremento de la altura de cada variedad en relación con el testigo correspondiente.

Variedad	E1	E2	E3
1	238.00	171.80	42.90
2	257.00	76.00	42.90
3	8.90	4.80	-38.70
4	-10.00	22.20	-38.70
5	51.30	63.00	6.30
6	110.00	58.80	6.30
7	8.50	3.90	-6.30
8	29.00	46.60	-6.30
9	30.30	26.20	16.80
10	42.70	2.30	16.80
11	-1.90	-1.00	0.00
12	33.20	-13.60	5.80
13	20.60	17.00	5.80
14	-13.70	-2.10	5.90
15	26.20	18.40	2.20
16	153.00	34.50	-45.00
17	264.60	90.40	-45.00
18	0.00	0.00	0.00

Un valor negativo en esta matriz significa que al aplicar el estimulador del crecimiento, el efecto fue negativo, es decir, inferior al testigo.

El objetivo es obtener un ordenamiento (*ranking*) de la muestra, que permita conocer los individuos que presentaron un mejor comportamiento así como detectar aquellos con un peor comportamiento. En otras palabras, se necesita saber cuál de las 18 variedades reaccionó de la mejor forma ante los tres estimuladores del crecimiento.

Para ello se define el individuo ficticio como el vector: (-13.70, -13.60, -45.00), correspondiente a los peores valores por carácter.

Para calcular la distancia de cada uno de los 18 individuos de la muestra a este individuo ficticio, se utiliza la I-Distancia de Ivanovic (Marta Flores y Endeljain, 1972) definida como:

$$D(i,s) = \sum_{j=1}^{p} X_{ij} - X_{sj} \prod_{u=1}^{j-1} (1 - r_{ju})$$

donde *rju* es el coeficiente de correlación lineal simple entre X_j y X_u.

En el uso de la I-Distancia, es necesario definir un orden de prioridad entre las variables: primeramente, se fija la más importante, quedando ordenado el resto de las variables a partir de la correlación que tengan con esta característica definida como la más importante.

En nuestro ejemplo, la característica más importante era *E1*, luego se ubicaba *E2* y finalmente *E3*, debido a que la matriz de correlaciones asociada a los datos fue:

	E1	E2	E3
E1	1.00		
E2	0.81	1.00	
E3	0.16	0.29	1.00

Al calcular el vector de distancia, los resultados fueron los siguientes:

	I-Distancia	*Ranking*
1	5.61	1
2	5.41	2
3	0.47	17
4	0.34	18
5	2.27	6
6	2.88	4
7	1.26	16
8	1.65	12
9	2.15	8
10	2.19	7
11	1.38	13
12	1.75	10
13	1.74	11
14	1.30	15
15	1.80	9
16	2.71	5
17	3.38	3
18	1.31	14

Este resultado permite hacer un reordenamiento en la muestra, a partir del hecho de que las variedades con mayor valor de I-Distancia, son las de mejor comportamiento en el conjunto de los tres estimuladores del crecimiento analizados.

En este caso, las variedades 1 y 2 fueron las que respondieron de la forma más positiva al efecto de los estimulantes del crecimiento, mientras que las variedades 3 y 4 fueron las de peor comportamiento, es decir, para ambos casos, el crecimiento se afectó ante el efecto de los supuestos estimuladores del crecimiento.

♦ ***Otras aplicaciones en la agricultura***

Moya *et al*. (1995) aplicaron la I-Distancia de Ivanovic para seleccionar de un grupo de 35 variedades de tomate las de mejor comportamiento, atendiendo a un conjunto de variables en ellas observadas. Los indicadores medidos fueron: rendimiento/planta (gramos), altura y diámetro del fruto (cm), número de frutos por racimo y días hasta la primera cosecha.

Los autores consideraron para el cálculo de la distancia, al rendimiento como variable más importante y ordenó el resto de los indicadores, de acuerdo con la correlación de estos con el rendimiento.

Como resultado, se obtuvo que las variedades que presentaban un mejor comportamiento, atendiendo al conjunto de caracteres evaluados, eran Moneymaker, Ontario 7620, Línea 16 y Patriot UF.

Miriam Alvarez (1982) utilizó dos grupos de variedades de una colección de piña para clasificarlas, atendiendo a una serie de características fenológicas y biomatológicas durante la primera cosecha. Para realizar el Análisis Discriminante, se consideró como variable más importante el peso promedio de un fruto con corona.

El el primer grupo, fueron consideradas 11 variables y 27 variedades de piña, mientras que en el segundo grupo fueron consideradas 19 variables y 49 variedades. En ambos casos, la variedad cabezona fue la de los mayores valores de la I-Distancia y, por tanto, la de mejor resultado atendiendo al conjunto de características evaluadas.

Por otra parte, Marta Alvarez (1987) utiliza otros criterios para seleccionar los mejores individuos dentro de una muestra, con la particularidad de que en su caso, no se calcula la distancia de los individuos al valor ficticio formado por las peores características. El método usado en este trabajo consiste en hacer una reordenación por cada variable, es decir, darle un ranqueo a cada individuo dentro de cada característica, para finalmente seleccionar aquellos cuya suma de rango entre todas las características sea menor.

Algunas consideraciones

Cuando se utiliza el Análisis Discriminante, con el objetivo de hacer un reordenamiento en la muestra, es decir, con el propósito de seleccionar los individuos de un mejor comportamiento, atendiendo a una serie de características en ellos evaluadas, es de suma importancia hacer un estudio previo de la naturaleza de los datos, para con ello formarse un criterio acerca de cuál medida de distancia estadística utilizar. Tomar una decisión correcta en este sentido, constituye la etapa más importante dentro del análisis; de ello depende en gran medida que los resultados finales tengan la mayor confiabilidad posible.

Seleccionar una u otra distancia depende fundamentalmente de las escalas de medición de las variables, así como de las estructuras de correlaciones en ellas existentes. Así, por ejemplo, si las variables se miden en diferentes escalas, es aconsejable hacer uso de una distancia que contemple en su fórmula la varianza de cada variable, para de esta forma asegurarse que las diferencias entre individuos para cada característica no tengan el mismo peso.

Si comparamos por ejemplo la distancia Euclideana y la de Frechet definidas como:

Distancia Euclideana:

$$Drs = \sqrt{\sum_{j=1}^{p} (x_{rj} - x_{sj})^2}\ 0$$

Distancia de Frechet:

$$Drs = \sum_{j=1}^{p} (x_{rj} - x_{sj}) / var(x_j)\ 0$$

se puede apreciar que la Euclideana no tiene en cuenta la diferencia de escalas entre las variables, siendo, por tanto, ineficiente su uso en problemas donde exístan diferencias entre escalas de medición. Este inconveniente queda resuelto con la Distancia de Frechet, en la que sí se tienen en cuenta estas diferencias al aparecer en su definición el concepto de varianza.

Ahora bien, existen otros casos en los que no es aconsejable el uso de las dos distancias mencionadas anteriormente, para establecer un correcto reordenamiento de la muestra. Nos referimos a la situación en que exista una estructura de correlación marcada entre las variables; para estos casos, si se utilizan cualesquiera de las dos distancias anteriores, se puede caer en el error de considerar en más de una ocasión las mismas diferencias entre individuos, debido a la alta correlación existente.

Precisamente, para salvar esta situación, es que se introducen algunas distancias que incorporan en su definición el concepto de correlación entre variables, como es el caso de la I-Distancia de Ivanovic, la cual no obstante puede tener el inconveniente que exíge tener definida una característica como la más importante.

Marta Flores y Endeljain (1972) plantean que en caso de que se tenga más de una variable como la más importante dentro de un análisis, es aconsejable aplicar la I-Distancia en más de una ocasión y finalmente tomar como medida a considerar, el valor promedio de las distancias.

De cualquier modo, se considera muy importante y necesario realizar un estudio previo de los datos para realizar una buena elección; esto es lo que asegura, en última instancia, una interpretación correcta de los resultados.

Refiriéndonos al Análisis Discriminante, podemos decir que es una técnica multivariada muy similar al Análisis Factorial Discriminante; en ambos casos, se parte de una matriz de datos particionada por filas, es decir, se tienen grupos de individuos formados *a priori*, sobre los cuales queremos conocer si existen o no diferencias en cuanto a su comportamiento en el conjunto de variables en ellos observadas.

Recomendamos el uso del Análisis Factorial Discriminante, cuando tengamos una cantidad considerable de variables, ya que en tal caso, con la aplicación de esta método se explican solamente aquellos ejes discriminantes que aportan información sobre diferencias entre grupos; es una técnica multivariada que construye nuevas variables o ejes factoriales, permitiendo la reducción de dimensionalidad del problema objeto de estudio.

Por otra parte, no recomendamos la aplicación del método cuando tengamos una cantidad considerable de grupos; en tal caso, es muy difícil lograr un por ciento alto de buena clasificación, no indicando ello necesariamente que no existan diferencias entre los grupos. Pudiera ocurrir que existan entre tantos grupos, algunos con características similares.

Cluster análisis. Clasificación automática

♦ ***Introducción***

Desde tiempos remotos, existe interés en clasificar personas y animales en grupos; así, por ejemplo, en el siglo XVIII, Linneo crea su clasificación del reino vegetal y animal, la cual aún tiene vigencia (Gladys Linares, Liliam Acosta y Viviam Sistach, 1986).

Los trabajos de clasificación más recientes han sido en la biología y reciben el nombre de taxonomía numérica, la cual en su inicio fue un arte y no una ciencia; sin embargo, gradualmente se han venido desarrollando técnicas más objetivas que llevan a considerar la taxonomía numérica como una ciencia.

Las técnicas de clasificación tienen como objetivo fundamental establecer grupos de individuos, siguiendo el criterio de unificar dentro de un mismo grupo a aquellos elementos de la muestra que tengan características similares, atendiendo al conjunto de variables en ellos observadas.

Es muy importante, en algunos casos, poder establecer un agrupamiento de los individuos, ya que en ocasiones, cuando resulte difícil manejar una cantidad considerablemente grande de datos, se puede aprovechar el hecho que dentro de un grupo hay elementos más o menos homogéneos y trabajar con un representante de cada grupo.

En los métodos de clasificación juega un rol importante el índice de similitud o distancia utilizada; es a partir de él que se van a calcular las distancias entre individuos o grupos de ellos, resultado que se utiliza finalmente para conformar los grupos: estarán dentro de un mismo conglomerado aquellos individuos con un valor de distancia "relativamente pequeño" entre ellos.

♦ ***Nociones del fundamento matemático***

Como en todo Análisis Multivariado, se parte de una matriz *X* de individuos-variables:

$$X=\begin{bmatrix} & \vdots & \\ \cdots & ij & \cdots \\ & \vdots & \end{bmatrix}_{nxp}$$

En este caso, en la matriz inicialmente no existe partición alguna ni por filas ni por columnas; el objetivo del método es formar una partición de esta matriz por filas, o lo que es lo mismo, construir grupos de individuos, atendiendo al criterio de que en un grupo o conglomerado, deben estar aquellos elementos de la muestra que tengan características similares en el conjunto de caracteres evaluados.

El primer paso es hacer una transformación de la matriz inicial *X* a una matriz cuadrada *D* de orden *n*, en la que va a aparecer en la posición *ij* la distancia entre el individuo *i* y el individuo *j*.

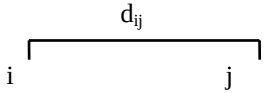
Para obtener la matriz *D*, se debe fijar de acuerdo con las características de los datos (escalas de medición de variables y estructuras de correlación), la distancia estadística o índice de similitud a utilizar. Una vez fijada la distancia a utilizar, se calcula la matriz *D*:

$$D=\begin{bmatrix} & \vdots & \\ \cdots & ij & \cdots \\ & \vdots & \end{bmatrix}_{n x n}$$

n: número de individuos

A partir de los datos de esta matriz, se comienza el proceso de formación de los grupos o conglomerados. Para ello exísten diferentes formas o procedimientos: en este material se considera la técnica jerárquica aglomerativa ascendente, para construir un dendograma o árbol de pertenencia, en la cual se sigue el siguiente esquema:

Primeramente se localiza en la matriz *D* el valor más pequeño, es decir, aquellos individuos más próximos entre sí: supongamos que este valor es d_{ij}; en tal caso se unen los individuos *i* y *j*, y se forma la base del árbol a partir de esta unión:

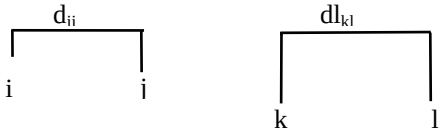


El próximo paso es formar una nueva matriz de distancias D_1, de orden *n-1*, en la que una fila y una columna va a representar como grupo a los individuos ya considerados; en este caso *i* y *j*.

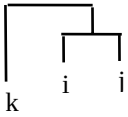
$$X=\begin{array}{c} 1 \\ 2 \\ \vdots \\ ij \\ \vdots \\ n-1 \end{array}\begin{bmatrix} & & \\ & \vdots & \\ \cdots & \vdots & \cdots \\ & \vdots & \end{bmatrix}_{(n-1)x(n-1)}$$

Lo único que cambia en esta matriz con relación a *D* es que en la fila y columna *ij* van a aparecer las distintas distancias de cada individuo al grupo formado por los individuos *i* y *j*; así, por ejemplo, en la posición 2, *ij* se va a encontrar el valor promedio entre d_{2i}, d_{2j} y d_{ij}.

El próximo paso es encontrar en D_1 el valor más pequeño: sea $d1_{kl}$ este valor, en tal caso se agrega una nueva capa al árbol o dendograma pero a un nivel más alto:



Si por el contrario, el valor buscado en *D1* hubiese sido un elemento de la fila *ij*, por ejemplo *ij*, *k* , entonces el árbol resultante sería:



que tiene la característica que en este caso el individuo *k* queda fusionado con el subgrupo *ij*.

El próximo paso es encontrar una nueva matriz D_2, y repetir el mismo procedimiento hasta que se llegue a una matriz de orden 2, o sea, hasta que queden incluidos todos los individuos o grupos de ellos.

Una vez finalizada la construcción del dendrograma o árbol de pertenencia, se llega a una decisión sobre cuáles van a ser finalmente los grupos formados; en este paso se tienen en cuenta los valores de distancia que ofrece el dendrograma, de forma tal de no considerar aquellas uniones con distancias "relativamente altas".

Por último, se puede hacer un estudio de la contribución de las variables en la formación de grupos o conglomerados, determinando en cada caso qué por ciento de la distancia entre dos individuos se atribuye a cada variable.

♦ ***Ejemplo de aplicación***

Para ejemplificar este método, se retoma el ejemplo utilizado en el epígrafe correspondiente al Análisis de Componentes Principales, en el cual se trabajan con las observaciones tomadas a 10 variedades de calabaza sometidas a ocho condiciones diferentes de estrés.

Matriz inicial de datos

		E1	*E2*	*E3*	*E4*	*E5*	*E6*	*E7*	*E8*
	1	8.44	8.01	2.50	3.29	1.33	0.80	1.84	1.19
	2	9.57	5.63	3.60	6.27	0.37	1.35	1.76	1.54
	3	10.93	5.42	3.70	6.10	0.90	0.99	1.30	0.14
	4	8.60	8.87	5.10	4.90	0.90	0.91	1.22	1.57
X=	5	9.60	3.37	1.45	2.80	0.35	0.56	0.97	0.62
	6	6.58	2.53	2.83	2.97	0.57	0.13	3.10	0.13
	7	4.66	3.64	2.80	1.52	0.35	0.32	2.74	0.37
	8	5.70	4.65	2.74	1.83	0.92	0.20	1.46	0.37
	9	5.33	3.03	1.92	2.91	1.09	0.27	1.71	0.54
	10	3.40	3.68	2.38	1.62	0.41	1.65	0.72	0.52

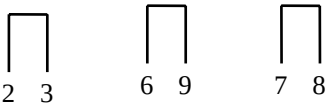
En este caso, se quiere hacer un reagrupamiento de las variedades de calabaza, atendiendo al criterio de incluir dentro de un mismo grupo o clase, aquellas que tengan un rendimiento más o menos similar en los ocho ambientes analizados.

Para dar cumplimiento al objetivo propuesto, se utiliza la distancia Euclideana como estimador de las proximidades entre variedades de calabaza o grupos de ellas.

Matriz de distancias entre variedades:

	Variedades	1	2	3	4	5	6	7	8	9
	2	4								
	3	5	2							
	4	3	4	5						
	5	5	5	5	7					
D=	6	6	6	6	8	4				
	7	6	7	8	8	6	3			
	8	5	6	7	7	4	3	2		
	9	6	6	7	8	4	2	2	3	
	10	7	8	9	9	6	5	3	3	3

En la primera etapa, puede apreciarse que los valores más pequeños de distancia se encuentran entre las variedades 2 y 3, entre las variedades 7 y 8, así como entre las variedades 6 y 9, todas ellas con un valor de distancia de dos unidades. Como consecuencia, en este primer corte se forman tres grupos o clases:

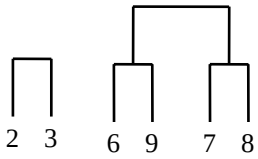


Constituyen por tanto estos tres grupos la base del dendograma, el cual se continúa formando a partir de calcular la matriz D_1, en la que aparece la distancia de cada uno de estos tres grupos al resto de las variedades, así como las distancias entre estas tres clases o grupos:

		1	2-3	4	5	6-9	7-8
	2-3	4.5					
	4	3	3.6				
D_1=	5	5	4	7			
	6-9	6	5.1	8	4		
	7-8	5.5	5.3	7.5	5	2.3	
	10	7	6.3	9	6	3.3	3

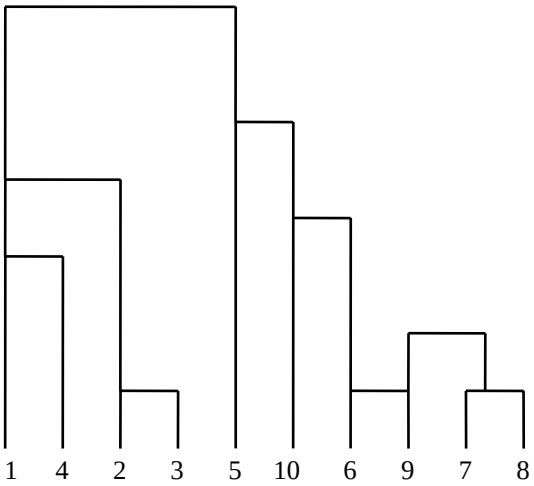
En esta matriz, el valor que aparece en la fila 6-9, columna 2-3, no es más que la distancia entre los grupos 6-9 y 2-3, formados en el paso anterior. Este valor se calcula como el promedio entre todas las distancias que se puedan establecer, en caso de concretarse esta unión; es decir, d_{69}, d_{62}, d_{63}, d_{92}, d_{93}, y d_{23}, los cuales aparecen en su totalidad en la matriz inicial de distancias D.

El próximo paso sería localizar en la matriz D_1, el valor más pequeño; en este caso 2.3, que representa la distancia entre los grupos 6-9 y 7-8. Esta información se incorpora al dendograma y, de esta forma, se avanza un nivel más en su construcción.



El siguiente paso sería calcular D_2, y todas las submatrices de distancia hasta llegar a enlazar a todos las variedades o grupos de ellas. Este procedimiento finalmente conduce al dendograma o árbol de pertenencia siguiente:

Dendograma



Para facilitar los pasos que siguen, se identifica cada nodo del dendograma con una enumeración:

Nodo 1 2-3
Nodo 2 7-8
Nodo 3 6-9
Nodo 4 6-9- 7-8
Nodo 5 1-4
Nodo 6 10- 6-9-7-8
Nodo 7 1-4- 2-3
Nodo 8 5- 10-6-9-7-8
Nodo 9 1-4-2-3- 5-10-6-9-7-8

Nótese que cada nodo representa los pasos que se siguieron en cada etapa para la construcción del dendograma.

Contribución de las variables:

Variedades	E1	E2	E3	E4	E5	E6	E7	E8
Nodo 1	41	1	0	1	6	3	5	43
Nodo 2	26	24	0	2	8	0	39	0
Nodo 3	31	5	17	0	5	0	38	3
Nodo 4	14	43	4	37	1	0	2	0
Nodo 5	0	7	62	24	2	0	4	1
Nodo 6	48	0	0	5	1	21	24	0
Nodo 7	18	51	0	26	1	1	0	2
Nodo 8	88	0	5	2	0	0	4	0
Nodo 9	35	35	5	23	0	1	0	1

Como puede apreciarse cada fila de esta matriz suma 100, y sus elementos por variable indican en qué por ciento contribuye cada variable en el cálculo de la distancia para la formación del nodo; así, por ejemplo, para el caso del nodo 1, el cual une las variedades de calabaza 2 y 3, la variable E3 no aporta nada en la diferenciación de estas variedades, en otras palabras, se puede decir que las variedades de calabaza 2 y 3 tuvieron un comportamiento similar en el ambiente 3.

Finalmente, una vez construido el dendograma, se está en condiciones de establecer los grupos siguiendo el criterio de hacer mínimas las diferencias dentro de un grupo.

Para la formación de los grupos existen varios criterios; en este trabajo se utiliza el que aparece en la programoteca francesa Statitcf en el módulo correspondiente al Cluster Análisis, el cual pide al usuario la cantidad de grupos o clases que desea formar.

En el ejemplo que nos ocupa se solicita la formación de tres grupos, los cuales quedaron constituidos como sigue:

Clase	Efectivo	Descripción
1	4	1,2,3,4
2	1	5
3	5	6,7,8,9,10

Valores medio por grupos y variables:

	E1	E2	E3	E4	E5	E6	E7	E8
Clase 1	9.4	7.0	3.7	5.1	0.9	1.0	1.5	1.1
Clase 2	9.6	3.4	1.5	2.8	0.3	0.6	1.0	0.6
Clase 3	5.1	3.5	2.5	2.2	0.7	0.7	1.7	0.7

Se puede concluir diciendo que el grupo 1, formado por las variedades de calabaza 1, 2, 3 y 4 fueron de forma general las que mejor se adaptaron a los ocho ambientes o condiciones de temperatura y humedad consideradas en el trabajo.

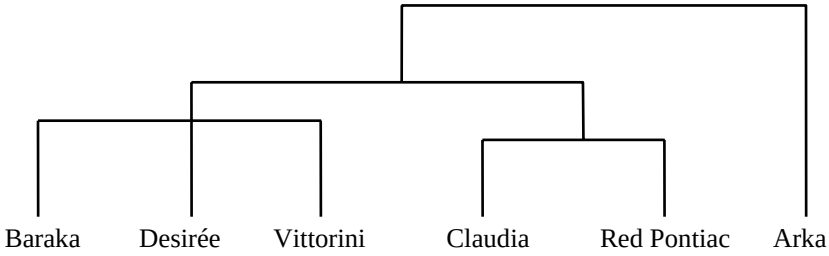
♦ ***Otras aplicaciones en la agricultura***

La mayor aplicación de los métodos de Cluster Análisis en la agricultura, lo es sin lugar a dudas en lo relacionado con trabajos encaminados a programas de selección y mejoramiento genético, en los cuales se pretende realizar cruzamientos entre variedades para lograr mejorar determinado carácter.

Arzuaga (1987) utilizó esta técnica para buscar estructuras de agrupación en seis variedades de papa, atendiendo al comportamiento de éstas en cinco evaluaciones semanales, en las que se evaluó la resistencia de las variedades a la *Alternaria solani*.

Las evaluaciones de severidad de ataque del hongo, se realizaron a partir del momento en que aparecieron los primeros síntomas. Para ello se utilizó una escala de nueve grados (1-9), según Horsfall y Barratt (1945).

En el trabajo se obtuvo el siguiente dendograma o árbol de pertenencia:



A partir del dendograma, el autor consideró la formación de tres grupos o cluster: un primer grupo formado por las variedades Baraka, Desirée y Vittorini, un segundo grupo formado por las variedades Claudia y Red Pontiac, y un tercer grupo formado por la variedad Arka, la cual ha sido informada como muy susceptible a la *Alternaria solani*.

En otro trabajo, pero en este caso en el cultivo del tomate, Le Minh Hong (1992) estudió la posibilidad de establecer estructuras de grupo a partir de 16 variedades, según el comportamiento de las mismas en una serie de indicadores evaluados.

Los caracteres analizados fueron: rendimiento, número de frutos por planta, masa promedio de los frutos, número de flores por racimo, porcentaje de fructificación, número de semillas por fruto, índice de cosecha, masa seca total, área foliar y masa específica de la hoja.

Para establecer los grupos, se construyeron 10 nuevas variables incorrelacionadas a partir del método propuesto por Rao (1952). Se calcularon los valores de distancia D^2 de Mahalanobis (Rao, 1952) entre todos los pares de variedades posibles.

El autor obtuvo con la aplicación del método, que en ambas épocas de estudio (primavera y verano), la variedad Nagcarlán fue la que más distó del resto.

El autor concluye, además, diciendo que los cruces de la variedad Campbell-28 con las variedades que más disten de ella, pudieran permitir la obtención de segregantes deseables para las condiciones adversas de siembra en cada época, atendiendo a que los cruces entre progenitores divergentes resultan, por lo general, idóneos para obtener segregantes o combinaciones deseables y producir un alto efecto heterótico y máxima variabilidad genética.

♦ ***Algunas consideraciones***

En el caso del Cluster Análisis, una vez que se requiere en su aplicación del cálculo de una distancia estadística para describir las proximidades entre individuos, se considera válido todo lo planteado en el epígrafe correspondiente al Análisis Discriminante, es decir, constituye la etapa más importante en la concepción del método, el momento en que se debe seleccionar una u otra distancia; una mala elección conduce a interpretaciones completamente sesgadas de un fenómeno objeto de investigación.

Una correcta elección en cuanto a la distancia a utilizar, para describir las proximidades entre individuos, es lo que va a permitir en última instancia eliminar diferencias en cuanto a diferentes escalas de medición de variables, a través de la utilización de distintas ponderaciones (Seber, 1984).

La elección de variables es otro aspecto importante a tener en cuenta. En tal sentido, es conveniente puntualizar que resulta sumamente económico hacer un estudio previo de las variables, para eliminar del análisis todas aquellas características que no ofrecen diferencias entre los individuos o tratamientos analizados; ello bien pudiera hacerse a través de las técnicas del Análisis de Varianza clásico.

Refiriéndonos al problema de la selección de variables, tampoco es conveniente considerar solamente características que ofrezcan un comportamiento demasiado diferenciado entre los individuos; de esta forma, resulta prácticamente imposible poder establecer cualquier reagrupamiento.

Análisis de varianza multivariado

♦ *Introducción*

Si bien en la Estadística univariada, el análisis de varianza tiene como objetivo esencial, determinar si un conjunto de tratamientos considerados en un experimento producen efecto estadísticamente diferenciado, atendiendo al comportamiento de los mismos en el carácter o variable analizada, en el Análisis de Varianza Multivariado (Manova), el objetivo es establecer si existen o no diferencias estadísticas entre los tratamientos, pero con la diferencia de que en este caso, interesa el comportamiento de los mismos en más de una variable o característica (Johnson, 1982).

Existen situaciones en las que cada parcela nos ofrece valores relativos a más de una variable; por ejemplo, en experimentos de asociación de cultivos, se pueden tener datos de rendimiento que bien pudieran estar referidos a los cultivos de frijol (X_1) y maíz (X_2), (Pimentel, 1987). En tal caso, una forma de proceder, si se quiere saber si el conjunto de tratamientos provoca un efecto diferenciado en los rendimientos de frijol y maíz en su conjunto, es realizar una transformación de los datos en ganancia, es decir, si sabemos que cada kilogramo de frijol vale cinco veces un kilogramo de maíz, podemos obtener una variable Y de ganancia como sigue:

$$Y= 5X_1 + X_2$$

y aplicar a los valores de Y un análisis de varianza univariado, aunque en este caso se ignoren estructuras de correlaciones entre las variables que bien pudieran influir en los resultados finales.

El Análisis de Varianza Multivariado permite dar solución a problemas como el anterior, en el que se quiere estudiar el efecto de un conjunto de tratamientos, atendiendo a su comportamiento en más de una variable.

♦ *Nociones del fundamento matemático*

Al igual que en el Análisis Factorial Discriminante, el fundamento matemático del método consiste en hacer uso del teorema de Fischer, para la descomposición de la suma de cuadrados del anova en componentes aditivas (Cooley y Lohnes, 1971).

Se parte de una matriz X de individuos-variables, particionada por filas, en donde cada partición corresponde a las distintas repeticiones de un tratamiento.

Para facilitar la explicación del método, se considera que el diseño estadístico utilizado es balanceado, o sea, que el número de repeticiones por tratamiento es el mismo (r).

Sea t la cantidad de tratamientos a probar, y p el número de variables analizadas; en tal caso, $n=t*r$ constituye el cantidad de observaciones totales por variable.

Si consideramos el elemento x_{ijk} $(i=1..t, j=1..p, k=1..r)$ de la matriz inicial X, como el valor que toma en la observación k, el tratamiento i en la variable j, se está en condiciones de calcular las correspondientes matrices de sumas de cuadrados y productos total, sumas de cuadrados y productos de tratamientos y sumas de cuadrados y productos de los residuos, como sigue:

Sean $A=(a_{ls})$, $H=(h_{ls})$ y $E=(e_{ls})$ las matrices mencionadas anteriormente en el orden respectivo en que aparecen, entonces:

$$a_{ls} = \sum_{i=1}^{t}\sum_{k=1}^{r} x_{ilk}\, x_{isk} - ((\sum_{i=1}^{t}\sum_{k=1}^{r} x_{ilk})(\sum_{i=1}^{t}\sum_{k=1}^{r} x_{isk}))/n$$

$$h_{ls} = (\sum_{i=1}^{t}(\sum_{k=1}^{r} x_{ilk} \sum_{k=1}^{r} x_{isk}))/r - (\sum_{i=1}^{t}\sum_{k=1}^{r} x_{ilk} \sum_{i=1}^{t}\sum_{k=1}^{r} x_{isk})/n$$

$$e_{ls} = a_{ls} - h_{ls}$$

Nótese que en este caso, al igual que en el análisis de varianza univariado, la parte correspondiente a los residuos se obtiene por diferencia entre la parte correspondiente al total y la parte correspondiente a las fuentes de variación controlables, sólo que en el caso multivariado el trabajo se efectúa con matrices.

Una vez calculada cada una de estas matrices, se recoge toda la información en la tabla de análisis de varianza multivariado:

Causas de variación	Grados de libertad	Matrices
Tratamientos	t-1	H= $\left[h_{1s}\right]$
Residuo	n-t-2	E= $\left[e_{1s}\right]$
Total	n-1	A= $\left[a_{1s}\right]$

El estadígrafo *F* multivariado, que se calcula en este caso para llegar a una conclusión sobre si existen o no diferencias entre los *t* tratamientos analizados, atendiendo a su comportamiento en el conjunto de *p* variables evaluadas, se basa en el criterio de Wilks, tratado por Rao (1952), el cual se representa con la letra griega lambda y se calcula como sigue:

lambda= det(E)/det(A), o lo que es lo mismo, se calcula a partir del cociente de dos determinantes.

Seguidamente a partir de este valor, se calcula la *F* multivariada como:

$$F = (k_2\text{-}1)/(k_1)*(1-\sqrt{\lambda})/(\sqrt{\lambda})$$

donde k_2 son los grados de libertad del residuo y k_1 representa los grados de libertad para tratamientos.

Este valor de *F* se compara con la *F* tabulada con $2k_1$ grados de libertad en el numerador y $p(k_2\text{-}1)$ grados de libertad en el denominador, para un nivel de confianza prefijado.

Si el valor *F* calculado excede al valor tabulado, se acepta la hipótesis de que los *t* tratamientos analizados, tienen un comportamiento diferenciado en el conjunto de los *p* caracteres considerados; en caso contrario, se rechaza la hipótesis de diferencia entre los tratamientos.

♦ ***Ejemplo de aplicación***

Para explicar en detalles la técnica de Análisis de Varianza Multivariado, tomaremos un ejemplo que aparece en Pimentel (1987); este consiste en probar el efecto de dos abonos orgánicos (turba fermentada y turba natural), en el proceso de asimilación por la planta del nitrógeno (N) y fósforo (P) del suelo.

El experimento se realizó siguiendo un diseño Completamente Aleatorizado, con cinco observaciones por tratamiento, los cuales a su vez consistieron en los dos abonos orgánicos mencionados anteriormente, más el testigo. Se consideraron las variables (X_1) como los valores de nitrógeno en la planta, y X_2 como los valores de fósforo en la planta.

Datos experimentales:

1- Testigo		2- Turba fermentada		3- Turba natural	
X_1	X_2	X_1	X_2	X_1	X_2
4.63	0.95	6.03	1.08	4.71	0.96
4.38	0.89	5.96	1.05	4.81	0.93
4.94	1.01	6.16	1.08	4.49	0.87
4.96	1.23	6.33	1.19	4.43	0.82
4.48	0.94	6.08	1.08	4.56	0.91

Las matrices *A*, *H* y *E* para este ejemplo son las siguientes:

$$A=\begin{bmatrix} 7.7056 & 0.9051 \\ 0.9051 & 0.1929 \end{bmatrix}$$

$$H=\begin{bmatrix} 7.2476 & 0.7326 \\ 0.7326 & 0.0982 \end{bmatrix}$$

$$E=\begin{bmatrix} 0.4580 & 0.1725 \\ 0.1725 & 0.1929 \end{bmatrix}$$

Fuente de variación	Grados de libertad	Matrices
Tratamientos	2	$H=\begin{bmatrix} 7.2476 & 0.7326 \\ 0.7326 & 0.0982 \end{bmatrix}$
Residuo	12	$E=\begin{bmatrix} 0.4580 & 0.1725 \\ 0.1725 & 0.1929 \end{bmatrix}$
Total	14	$A=\begin{bmatrix} 7.7056 & 0.9051 \\ 0.9051 & 0.1929 \end{bmatrix}$

En el ejemplo que se considera:
det(E)=0.013616 y det(A)=0.667204, por lo que *lambda= 0.013616/0.667204= 0.0204*, lo cual equivale a un valor de *F* igual a:

$$F = \frac{12-1}{2} * \frac{1-\sqrt{0.0204}}{\sqrt{0.0204}} = 33.01$$

con 2*2=4 y 2*(12-1)=22 grados de libertad para el numerador y denominador respectivamente.

Para un nivel de confianza del 95 %, la *F* tabulada con 4 y 22 grados de libertad es 2.82.

Como el valor calculado es mayor que el tabulado, se puede afirmar que exísten diferencias entre los dos abonos orgánicos y el testigo, atendiendo a los niveles de nitrógeno y fósforo que ellos provocan en la planta.

Correlaciones canónicas

♦ ***Introducción***

El Análisis de Correlaciones Canónicas es un método multivariado que permite investigar acerca de la existencia de correlación entre dos grupos de variables.

En una investigación en la rama agrícola, podemos estar interesados en determinar si existe influencia de un grupo de variables independientes sobre otro conjunto de variables dependientes. Así, por ejemplo, puede ser de interés conocer acerca de la influencia de un conjunto de variables climáticas sobre un conjunto de variables componentes del rendimiento.

Con la aplicación de esta técnica multivariada de análisis de datos, no solamente damos respuesta a la interrogante anterior; además de ello, podemos determinar en caso de existir asociación, las variables responsables dentro de cada grupo de tal relación.

♦ ***Nociones del fundamento matemático***

En este caso, como en todo método multivariado, se parte de la matriz inicial de datos:

$$X=\begin{bmatrix} & \vdots & \\ \cdots & ij & \cdots \\ & \vdots & \end{bmatrix}_{nxp}$$

Así la matriz de datos X aparece particionada por columna; cada partición representa un conjunto de variables asociadas por determinada característica común.

El método consiste en encontrar pares de nuevas variables canónicas, formadas por una combinación lineal de las variables originales de cada grupo, de forma tal que la correlación entre estas componentes canónicas sea máxima (Kendall, 1980).

Sean: $X_1,X_2,...,X_k$ las variables correspondientes al primer grupo

$X_{k+1},X_{k+2},...,X_p$ las variables correspondientes al segundo grupo

Supongamos, además, sin perder generalidad que $k <= p-k$, es decir, el número de variables del primer grupo es menor o igual al número de variables del segundo grupo.

Se obtienen tantos pares de variables canónicas como variables tenga el grupo más pequeño. En nuestro caso, se pueden formar *k* pares de variables canónicas.

Se construyen entonces las dos primeras variables canónicas u_1 y v_1 como una combinación lineal de las variables del primer y segundo grupos respectivamente:

$$U_1= \alpha_{11}X_1+.....+\alpha_{1k}X_k \qquad V_1= \alpha_{1k+1}X_{k+1}+.....+\alpha_{1p}X_p$$

Como se dijo anteriormente, los coeficientes α_{ij} se buscan de tal forma que la correlación lineal entre U_1 y V_1 sea máxima.

Ello se traduce en encontrar los valores y vectores propios de la matriz:

$$M_{kxk}= C_{12}C_{22}^{-1}C_{21}$$

Las submatrices C_{ij} *i=1..2, j=1..2* se extraen de la matriz de varianzas y covarianzas de *X*:

$$\text{cov}(X)=\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

es decir:

C_{11} representa la submatriz de varianzas y covarianzas entre las variables del primer grupo

C_{12} representa la submatriz de varianzas y covarianzas entre las variables del primer grupo con las variables del segundo grupo

C_{22} representa la submatriz de varianzas y covarianzas entre las variables del segundo grupo

Sea λ_1^2 el mayor valor propio de la matriz *M*. El vector propio asociado a λ_1^2, coincide con el vector $(\alpha_{11},.....,\alpha_{1k})$, es decir, con los coeficientes necesarios para la formación de U_1.

La otra componente canónica V_1 se busca como la proyección ortogonal de u_1 sobre R^{p-k}.

La correlación entre U_1 y V_1 no es más que λ_1.

El segundo par de variables canónicas, U_2 y V_2, debe explicar la mayor parte posible de la asociación entre los dos grupos de variables, que no fue explicada por el primer par de variables canónicas U_1 y V_1. Esto es equivalente a decir que U_2 y V_2 tienen correlación máxima sujeta a la condición:

$$Cor(U_1\ U_2)=0 \qquad Cor(V_1\ V_2)=0$$

Se demuestra que los coeficientes para la construcción de U_2 no son más que el vector propio asociado al segundo mayor valor propio de M (λ_2^2). En este caso, λ_2 representa la correlación entre U_2 y V_2. Se cumple además que $\lambda_1 \geq \lambda_2$.

De igual forma, V_2 se encuentra como la proyección ortogonal de U_2 sobre R^{p-k}.

Finalmente se sigue el mismo algorítmo hasta encontrar los *K* pares de variables canónicas.

A pesar de que con la aplicación del Análisis de Correlaciones Canónicas se construyen *k* pares de variables canónicas, generalmente no resulta necesario explicarlos en su totalidad.

¿Cómo saber con cuántos pares de variables canónicas trabajar?

Para dar respuesta a la interrogante anterior, lo primero que se hace es la prueba de hipótesis siguiente:

$$H_0: \lambda_1=\lambda_2=.....=\lambda_k=0 \qquad H_A: \exists_i \ \lambda_i \neq 0$$

Para llegar a una regla de decisión, se construye el estadígrafo:

$$U_1 = -[n-\tfrac{1}{2}(p+1)]\ln(1-\lambda_1^2)$$

Se demuestra que *U* distribuye Chi-Cuadrado con *(k-1)*(p-k-1)* grados de libertad.

Regla de decisión: aceptar H_0 si $U \leq X^2_{(1-\alpha)}\ [(k)*(p-k)]$

En caso contrario, se rechaza $H_{0,}$ es decir, se acepta que existe al menos un $\lambda_i \neq 0$.

Ahora bien este $\lambda_i \neq 0$ tiene que ser λ_1 , ya que como se sabe:

$$\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_k$$

Por otra parte, decir que λ_1 es estadísticamente diferente de cero, es equivalente a decir que la correlación entre U_1 y V_1 es significativa, o lo que es lo mismo, significa que existe relación de dependencia entre los dos grupos de variables, resultando necesario explicar al menos este primer par de variables canónicas.

En caso de haber aceptado $H_{0,}$ significaría que todos los λ_i son estadísticamente iguales a cero y, por tanto, se acepta que no existe relación entre los dos grupos de variables; resultando innecesario verificar la posible inclusión del resto de los pares de variables canónicas.

Si por el contrario se rechaza H_0, cabe la posibilidad de que no sea suficiente con el primer par U_1, V_1 para explicar la relación de asociación entre los dos conjuntos de variables:

¿Cómo saber si resulta necesario explicar el próximo par U_2 ,V_2?

Para ello se calcula el estadígrafo:

$$U_2 = -[n-\tfrac{1}{2}(p+1)]\ln(1-\lambda_2^2)$$

U_2 se compara con $X^2_{(1-\alpha)}$ con $(k-1)(p-k-1)$ grados de libertad.

Si $U_2 \geq X^2[(k-2)*(p-k-2)]$ se rechaza, $H_{0,}$ o lo que es lo mismo, se incluye el segundo par de variables canónicas U_2 y V_2, y se analiza por el mismo algoritmo, si resulta necesario trabajar con un tercer par de variables canónicas.

En caso contrario, es decir, si aceptamos H_0, se concluye que el segundo par de variables canónicas no es necesario explicarlo, y que solamente es necesario trabajar con un solo par de variables canónicas.

Lo anterior es debido a que si $\lambda_2=0$, los λ_i siguientes han de ser también estadísticamente iguales a cero.

Una vez decidido con cuantos pares de variables canónicas trabajar, resulta necesario dar una explicación a las variables canónicas a partir de las variables originales. Para ello, se seleccionan en cada combinación lineal aquellos coeficientes con mayor valor absoluto. Las variables por cada grupo asociadas a los coeficientes seleccionados, serán las responsables de la relación entre los grupos de variables.

♦ *Ejemplo de aplicación*

Para ejemplificar el uso del método, utilizaremos un juego de datos que no está vinculado a ninguna situación real.

Supongamos que queremos investigar acerca de la relación entre tres caracteres climáticos (X_1, X_2 y X_3) sobre cuatro caracteres agronómicos (Y_1, Y_2, Y_3, Y_4), en 14 variedades de arroz.

Datos experimentales:

Variedades	X_1	X_2	X_3	Y_1	Y_2	Y_3	Y_4
1	30	114	56	26.4	32.1	58.0	0.70
2	40	112	62	26.0	30.0	78.0	0.58
3	50	100	50	30.1	33.4	99.4	0.60
4	60	98	44	32.0	35.0	118.1	0.40
5	30	94	44	33.4	38.2	58.9	0.32
6	35	90	36	35.0	31.0	68.0	0.41
7	40	86	38	39.2	34.3	79.8	0.62
8	45	120	61	20.0	33.0	93.0	0.34
9	20	124	62	18.6	30.0	38.6	0.38
10	25	100	48	31.0	31.2	51.0	0.40
11	20	100	49	32.3	31.2	42.0	0.40
12	20	114	50	27.0	30.0	43.0	0.61
13	35	100	50	29.8	30.0	68.0	0.62
14	35	100	71	28.9	28.4	69.3	0.70

Resultados:

Con la aplicación del método se forman tres pares de variables canónicas:

Vectores propios (Coeficientes de las variables canónicas)

Variable	primer par canónico	segundo par canónico	tercer par canónico
X_1	**-1.026**	-0.104	0.135
X_2	-0.128	**-0.834**	1.169
X_3	-0.023	-0.239	-1.383
Y_1	0.136	**1.027**	-0.295
Y_2	-0.013	-0.055	0.723
Y_3	**-1.007**	-0.009	-0.212
Y_4	-0.038	-0.167	-0.509

A partir de la tabla anterior, se forman los pares de variables canónicas:

$U_1 = -1.026X_1-0.128X_2-0.023X_3$

$V_1 = 0.136Y_1-0.013Y_2-1.007Y_3-0.038Y_4$

$U_2 = -0.104X_1-0.834X_2-0.239X_3$

$V_2 = 1.027Y_1-0.055Y_2-0.009Y_3-0.167Y_4$

$U_3 = 0.135X_1+1.169X_2-1.383X_3$

$V_3 = -0.295Y_1+0.723Y_2-0.212Y_3-0.509Y_4$

Prueba de significación para los pares de variables canónicas:

Par	λ_i^2	X^2	g.l	Significación
U_1 V_1	0.995	82.15	12	0.00001
U_2 V_2	0.971	34.39	6	0.0002
U_3 V_3	0.224	2.29	2	0.317

De los tres pares de variables canónicas que se formaron, solamente es necesario explicar las dos primeras, con coeficientes de correlación:

$corr(U_1 V_1) = \lambda_1 = \sqrt{0.995} = 0.9974$

$corr(U_2 V_2) = \lambda_2 = \sqrt{0.971} = 0.9850$

Al ser λ_1 estadísticamente diferente de cero, podemos decir que existe una relación de dependencia entre los dos grupos de variables, es decir, en la muestra de las 14 variedades de arroz estudiadas, las variables climáticas están influyendo en las variables agronómicas.

En el primer par de variables canónicas, se aprecia una estrecha relación entre X_1 y Y_3, siendo una relación directamente proporcional. Por otra parte, con el segundo par de variables canónicas, se observa el efecto que hace X_2 en Y_1 ; en este caso, a medida que aumenta X_2 disminuye Y_1.

♦ ***Otras aplicaciones en la agricultura***

El Análisis de Correlaciones Canónicas fue utilizado por Lourdes Iglesias (1986), la cual en su tesis de Doctorado aplicó este método para investigar sobre la existencia de una posible asociación entre cinco caracteres morfológicos y siete caracteres agronómicos, en dos épocas de siembra, en el cultivo de la soya.

Como resultado de la aplicación del método, se encontró una correlación altamente significativa (0.7201) en verano, no ocurriendo así en la época de invierno, donde la correlación no fue significativa.

En la época de verano se materializó la alta correlación, debido fundamentalmente a la relación existente entre el rendimiento y el número de vainas por planta por la parte agronómica, con el número de nudos fértiles por la parte morfológica.

Análisis factorial de correspondencia

El Análisis Factorial de las Correspondencias Múltiples es un método análogo al Análisis de Componentes Principales, pero en este caso diseñado para tratar con variables de naturaleza discreta.

Al igual que en las Componentes Principales, se construyen nuevas variables, pero con la peculiaridad de que se construyen como combinaciones lineales de las categorías de las variables iniciales. Los coeficientes para la formación de las nuevas variables o componentes, se calculan a partir de la diagonalización de matrices de perfiles o tablas de contingencia.

Entre las ventajas del método, se puede citar la de permitir hacer una representación gráfica en un mismo plano de los individuos y las categorías de las variables, lo cual permite estudiar el comportamiento de los individuos por su proximidad con determinada categoría dentro de una variable.

Este método fue utilizado por Lourdes Iglesias (1987), con el objetivo de conocer el grado de divergencia bioquímica existente entre 18 variedades de soya, teniendo en cuenta el comportamiento de ellas en ocho variables de naturaleza discreta:

- Contenido proteíco del grano (tres categorías)
- Variantes peroxidasas semilla (dos categorías)
- Variante peroxidasa raíz (dos categorías)
- Variante amilasa semilla (dos categorías)
- Variante amilasa hojas (dos categorías)
- Variante catalasa semilla (dos categorías)
- Variante proteína semilla (dos categorías).

Como resultado de la aplicación del método, se establecieron seis grupos de variedades: tres de ellos constituidos por una sola variedad, destacándose además la notable diversidad bioquímica existente en el grupo representado por la variedad Vavilov 6317, cuya procedencia, grupo de maduración y características morfoagronómicas distintivas la alejaron sustancialmente del conjunto varietal examinado.

En otro trabajo, Mayra E. González (1991) utilizó el Análisis Factorial de Correspondencia en datos provenientes de una encuesta agrícola, en donde se tomaron como individuos una serie de campos con diferentes características, entre las que se destacaban la diferencia de edades de ellos.

Como resultado de la aplicación del método, la autora detectó la exístencia de una asociación entre los valores más bajos de aplicación y asimilación de los nutrientes, N, P, K y las clases de valores más pequeños del rendimiento; la asociación se mantuvo para los niveles intermedios, pero en el caso del nitrógeno, se encontró que las aplicaciones más altas no están asociadas precisamente a los mayores rendimientos.

Referencias

Alvarez, Miriam. Una aplicación del método de la I-Distancia a la selección de grupos de variedades de piña (*Ananas comosus* L. Merr). Cultivos Tropicales 4(3):427-435, 1982.

Alvarez, Miriam y María M. Hernández. Estudio de los Componentes Principales en un grupo de variedades de plátano. Cultivos Tropicales 4(2):227-240, 1982.

Alvarez, Marta. Mejoramiento genético del tomate *Lycopersicon esculentum* Mill para siembras de primavera /Marta Alvarez Gil.- Tesis de grado (Dr. en Ciencias Agrícolas), INCA, 1987.- 91h.

Anderson, J. An Introduction to multivariate statistical analysis. /J. Anderson.- New York: John Wiley and Sons, 1968.

Arzuaga, J. Clasificación de seis variedades de papa por su resistencia a *Alternaria solani* mediante un método de clasificación automática. Cultivos Tropicales 9(2): 59-63, 1987.

Cooley, W. W. Multivariate Data Analysis. /W. W. Cooley, P.R. Lohnes.- New York: John Wiley and Sons, 1971.- 563p.

Cruz, R. /*et al*./. Evaluación de progenitores de caña de azúcar (*Saccharum* spp) del programa de mejoramiento genético, con fines comerciales de las provincias orientales de Cuba. Cultivos Tropicales 16(2):70-73, 1995.

Cuthbert, D. Fitting equations to data computer analysis of multifactor data. /D. Cuthbert and F. S. Wood. /2da. edición. New York/ John Wiley and Sons, 1980.- 458 p.

Dell'Amico, J. M. Comportamiento de plantas de tomate (*Lycopersicon esculentum* Mill) ante diferentes condiciones de abastecimiento hídrico del suelo.- Tesis de grado (Dr. en Ciencias Agrícolas); INCA, 1992.-168 h.

Dempster, A. Elements of continuous multivariate analysis. /Massachussetts, Addison Wesley, 1969.- 288p.

Flores, Marta. Análisis Discriminante. /Marta Flores, V.Endeljain.- La Habana : Universidad de La Habana, 1972.- 60 p.

Gnanadesikan, R. Statistical data analysis of multivariate observations. /R. Gnanadesikan.- New York : John Wiley and Sons, 1977.- 311 p.

González, María C. /*et al*./. Análisis de la variabilidad originada por el cultivo *in vitro* de semillas de la variedad Amistad-82 en condiciones salinas. Cultivos Tropicales 12(3): 83-85, 1991.

González, Mayra E. Análisis de datos en la interpretación del rendimiento en una encuesta agrícola. /Mayra E. González.- Tesis de grado (Dr. en Ciencias Agrícolas); ISACA, 1991.- 145 p.

Hope, K. Methods of multivariate analysis. K. Hope/ London/ 1968.- 165 p.

Iglesias, Lourdes. Estudio del grado de divergencia bioquímica en un grupo de variedades de soya mediante el empleo del Análisis Factorial de Correspondencia. Cultivos Tropicales 9(2):47-54, 1987.

Iglesias, Lourdes. Estudio de la variabilidad morfoagronómica y bioquímica en soya. /Lourdes Iglesias.- Tesis de grado (Dr. en Ciencias Agrícolas); INCA, 1986.- 232 p.

Iglesias, L. A y Lourdes Iglesias. Clasificación del comportamiento de variedades de trigo en Cuba mediante el método de Análisis de Componentes Principales. Cultivos Tropicales 16(2):66-69, 1995.

Johnson, R. A. Applied Multivariate Statistical Analysis. /Richard. A. Johnson and Dean. W. Wichern..- Prentice-Hall, Inc, Englewood Cliffs, New Jersey.- University of Winconsin, 1982.- 591 p.

Judez, L. Técnicas de análisis de datos multidimensionales. /L. Judez.- Ministerio de Agricultura, Pesca y Alimentación. Secretaría General Técnica. Madrid, 1989.- 301 p.

Kendall, M. Multivariate Analysis. /M. Kendall.- Charles Griffin and Company Ltt, 1980.- 209 p.

Krzonowski, W. J. Principles of Multivariate Analysis. /W. J Krzonowski. Tomo I./ Sever editors.- J. B. Copas/Deparment of Applied Statistical University of Reading.- Oxford, 1988.- 284 p.

Le Minh Hong. Selección de progenitores para el mejoramiento genético del tomate (*Lycopersicon esculentum* Mill) en siembras fuera de época. /Le Minh Hong.- Tesis de grado (Dr. en Ciencias Agrícolas); INCA, 1992.- 148 h.

Linares, Gladys. Estadística Multivariada. /Gladys Linares, Liliam Acosta, Viviam Sistach.- La Habana : Universidad de La Habana, 1986.- 319 p.

Morrison, D. F. Multivariate Statistics. /D. F. Morrison.- New York : John Wiley and sons, 1979.- 414 p.

Moya, C. /*et al*./. Selección de progenitores, estimados de repetibilidad y correlaciones en tomate (*Lycopersicon esculentum* Mill) en condiciones de organopónico. Cultivos Tropicales 16(2):79-83, 1995.

Panse, W. G. Statistical methods for agricultural workers. /W.G Panse, P.V.Sukhatme.- 2.ed.-New Delhi : Indian Council of Agricultural Research, 1967.- 381 p.

Pérez, Noraida, C. Ismail y María .C. González. Mejoramiento genético mediante el cultivo *in vitro* de anteras de híbridos de arroz. Cultivos Tropicales 16(2):54-56, 1995.

Pimentel, F. A. Estadística moderne na pesquise agropecuarie. /F. A. Pimentel.- Piracicaba : Associacao Brasileira para Pesquisa de Potassa, 1987.- 162 p.

Plana, R. Caracterización del material de plantación en caña de azúcar (*Saccharum* sp) en un suelo Ferralítico Rojo compactado. Cultivos Tropicales 12(1): 35-39, 1991.

Rao, C. R. Advanced statistical methods in Biometrical research. /C. R. Rao.- New York : John Wiley and Sons, 1952.- 390 p.

Seber, G. A. F. Multivariate observations. G. A. F. Seber.- New York : John Wiley and Sons, 1984.- 670 p.

Este documento está dirigido a proporcionar las herramientas y conocimientos necesarios para acometer el trabajo de análisis e interpretación de los resultados donde intervengan mediciones múltiples efectuadas sobre un conjunto de individuos. La obra la conforma una recopilación de las principales técnicas de análisis de datos, tratadas con todo el rigor necesario en cuanto a su fundamentación teórica de cada una, las exigencias para su implementación, el alcance y las posibilidades de extraer información a los resultados, así como las limitaciones para su uso y aplicación. Cada una de estas técnicas se presentan con el desarrollo de una investigación real conducida en el instituto y analizada por el departamento. Este material en manos del investigador puede representar una importante guía para el análisis multivariado de datos, en tanto que en el aspecto docente constituye el documento base para los cursos de posgrado y el de cursos opcionales de Análisis Multivariado para las Maestrías en Ciencias Agrícolas y Veterinarias.

ISBN 950-7023-04-0

9 789597 023043