

ÍNDICE

PRÓLOGO	19
---------------	----

UNIDAD DIDÁCTICA V
EL ANÁLISIS MULTIVARIANTE
EN LA INVESTIGACIÓN EDUCATIVA

Capítulo 1
INTRODUCCIÓN AL ANÁLISIS MULTIVARIANTE

1. Introducción.....	27
2. La estadística y los métodos multivariantes. Breve recorrido histórico	27
3. Concepto de análisis multivariante.....	30
4. Tipos de variables y clasificación de los métodos multivariantes	32
5. Presente y futuro del análisis multivariante	37
Ejercicios de autocomprobación.....	40
Solución a los ejercicios de autocomprobación.....	41
Bibliografía.....	42

Capítulo 2
HERRAMIENTAS DEL ANÁLISIS MULTIVARIANTE

1. Herramientas algebraicas básicas.....	45
1.1. Matrices.....	45

1.2. Determinantes	48
1.3. Vectores	51
2. Herramientas estadísticas	57
2.1. La matriz de datos	57
2.2. Estadísticos	58
2.3. Distribuciones estadísticas	60
3. Herramientas informáticas: SPSS	65
3.1. Ventanas del SPSS	66
3.2. Menús	67
3.3. Barra de herramientas y barra de estado	68
3.4. Entrada y ejecución del SPSS	68
3.5. Pasos básicos para realizar un análisis de datos con SPSS.....	69
Ejercicios de autocomprobación.....	70
Solución a los ejercicios de autocomprobación.....	71
Bibliografía	74

Capítulo 3

LA SIMULACIÓN Y LA EXPLORACIÓN INICIAL DE DATOS

1. Introducción.....	77
2. Simulación y modelos de simulación.....	77
2.1. Creación de modelos de simulación	81
2.2. Modelos probabilísticos	83
3. Exploración inicial de los datos	88
3.1. Casos aislados y ausentes.....	88
3.2. Supuestos paramétricos de los datos	90
Ejercicios de autocomprobación.....	94
Solución a los ejercicios de autocomprobación.....	95
Bibliografía	96

Capítulo 4

ANÁLISIS FACTORIAL

1. Definición y objetivos	99
1.1. Definición	99
1.2. Objetivos	99
2. Diseño y método	100
2.1. Diseño	101
2.2. Modelo matemático	101
2.3. Teorema de Thurstone	103

3.	Supuestos y limitaciones	104
3.1.	Evaluación de la matriz de correlaciones	104
3.2.	Limitaciones	106
4.	Procedimiento	106
4.1.	Pasos para la realización del análisis factorial exploratorio ...	106
4.2.	Extracción de los factores	107
4.3.	Rotación de los factores	108
4.4.	Cálculo de las puntuaciones factoriales	111
5.	Interpretación de resultados	111
6.	Análisis factorial confirmatorio	113
7.	Un ejemplo de análisis factorial con SPSS	113
7.1.	Problema-ejemplo	113
7.2.	Procedimiento	113
7.3.	Presentación e interpretación de resultados	125
7.4.	Principales conclusiones	134
8.	Otro ejemplo de análisis factorial (componentes principales) y su resolución manual	134
	Ejercicios de autocomprobación.....	139
	Solución a los ejercicios de autocomprobación.....	149
	Bibliografía	151

Capítulo 5
LA TIPOLOGÍA Y EL ANÁLISIS DE CLUSTER

1.	Definición y objetivos	155
1.1.	Definición	155
1.2.	Objetivos	156
2.	Diseño y método	157
2.1.	Diseño	157
3.	Supuestos y limitaciones	177
4.	Procedimiento	178
4.1.	Métodos jerárquicos	178
4.2.	Métodos no jerárquicos	194
5.	Interpretación de resultados	195
6.	El análisis de Cluster mediante SPSS	197
6.1.	Introducción	197
6.2.	Análisis de conglomerados jerárquicos	197
6.3.	Análisis de conglomerados de k-medias	201

7. Ejemplos de análisis de Cluster con SPSS.....	204
7.1. Un ejemplo de clasificación jerárquica.....	204
7.2. Ejemplo de clasificación mediante el procedimiento no jerárquico de k-medias	212
Ejercicios de autocomprobación.....	221
Solución a los ejercicios de autocomprobación.....	224
Bibliografía.....	235

Capítulo 6

ANÁLISIS DE CORRESPONDENCIAS

1. Análisis de correspondencias simples	239
1.1. Definición y objetivos del método	239
1.2. Diseño y método	240
1.3. Supuestos y limitaciones	247
1.4. Procedimiento	247
1.5. Interpretación de resultados	250
1.6. Resolución informática mediante SPSS de un ejemplo.....	254
2. Análisis de correspondencias múltiples	266
2.1. Definición y objetivos	266
2.2. Diseño y método	266
2.3. Supuestos y limitaciones	273
2.4. Interpretación de resultados	273
2.5. El análisis de correspondencias múltiple: resolución informá- tica mediante SPSS de un ejemplo	274
Ejercicios de autocomprobación.....	283
Solución a los ejercicios de autocomprobación.....	289
Bibliografía.....	294

Capítulo 7

ANÁLISIS DISCRIMINANTE

1. Definición y objetivos	297
2. Diseño y método	299
2.1. Modelo matemático	299
2.2. Tamaño de la muestra	301
2.3. División de la muestra	301
3. Supuestos y limitaciones	301
3.1. Supuestos del modelo.....	301
3.2. Comprobación de los supuestos paramétricos	302
4. Procedimiento	303

4.1. Fases del análisis:	303
5. Interpretación de resultados	309
6. Un ejemplo de análisis discriminante con SPSS	310
6.1. Problema-ejemplo	310
6.2. Procedimiento	311
6.3. Presentación e interpretación de resultados	323
6.4. Principales conclusiones	335
7. Otro ejemplo de análisis discriminante para tres o más grupos	336
Ejercicios de auto comprobación.....	348
Solución a los ejercicios de auto comprobación.....	350
Bibliografía.....	358

Capítulo 8

ANÁLISIS DE REGRESIÓN

1. Definición y objetivos	361
2. Diseño y método	363
2.1. Diseño de investigación	363
3. Supuestos y limitaciones	365
4. Procedimiento	370
5. Interpretación de resultados	377
6. El análisis de regresión con SPSS. un ejemplo de aplicación	378
6.1. Problema ejemplo	378
6.2. Procedimiento	378
6.3. La elaboración de pronósticos	403
7. Análisis de regresión lineal múltiple. Un ejemplo de aplicación: solución matricial	404
Ejercicios de auto comprobación.....	409
Solución a los ejercicios de auto comprobación.....	414
Bibliografía.....	423

Capítulo 9

ANÁLISIS DE SEGMENTACIÓN. EL MODELO CHAID

1. Presentación	427
2. El análisis de segmentación. El modelo Chaid	428
2.1. Definiciones.....	428
2.2. Objetivos y aplicaciones del método	430

2.3. Relación con otros métodos multivariantes	432
2.4. El estadístico ji-cuadrado en la segmentación	433
2.5. El procedimiento Chaid	434
2.6. Un ejemplo de aplicación del procedimiento Chaid	437
2.7. Supuestos y limitaciones	440
2.8. Interpretación de resultados	440
3. El programa Answer Tree y el procedimiento Chaid	441
3.1. Introducción	441
3.2. El procedimiento Chaid en el programa Answer Tree: proceso de creación de un árbol	442
4. Un ejemplo de aplicación del Answer Tree a la investigación edu- cativa.....	451
4.1. Objeto del estudio: El clima laboral	451
4.2. Los datos	452
4.3. Objetivo y proceso	452
4.4. Desarrollo del análisis y creación del primer árbol	454
4.5. Principales conclusiones	459
Ejercicios de autocomprobación.....	462
Solución a los ejercicios de autocomprobación.....	465
Bibliografía.....	469

Capítulo 1
INTRODUCCIÓN
AL ANÁLISIS MULTIVARIANTE

ESQUEMA

1. *Introducción*
2. *La estadística y los métodos multivariantes. Breve recorrido histórico*
3. *Concepto de análisis multivariante*
4. *Tipos de variables y clasificación de los métodos multivariantes*
5. *Presente y futuro del análisis multivariante*

Ejercicios de auto comprobación

Solución a los ejercicios de auto comprobación

Bibliografía

OBJETIVOS

- Configurar un esquema claro del proceso evolutivo de los métodos de análisis multivariante, identificando aquellos hitos de mayor influencia en la Investigación Educativa
- Distinguir las aplicaciones de los métodos de análisis multivariante en el ámbito de la Investigación Educativa
- Valorar las diferentes clasificaciones de los métodos de análisis multivariantes, distinguiendo los elementos diferenciadores de las mismas

1. INTRODUCCIÓN

El creciente uso de las tecnologías de la información y de la comunicación ha facilitado los procesos de recogida, almacenamiento y envío de grandes —en volumen y complejidad— bases de datos procedentes de la experimentación o de la observación. Esta realidad ha demandado medios para el análisis acordes con la misma; de ahí el espectacular avance experimentado por los **métodos estadísticos multivariantes**

El empleo del análisis multivariante en la investigación educativa se enmarca dentro del enfoque cuantitativo de investigación, y en general, en diseños no experimentales, aún cuando se pueda utilizar en los estudios cualitativos, siempre que exista una cuantificación, y en diseños experimentales —dentro del denominado modelo lineal factorial.

Indudablemente, los métodos de análisis multivariante necesitan la confluencia de dos materias instrumentales: la Estadística y la Informática; pero la importancia de estas dos materias, para la investigación educativa, no es la misma. Históricamente, la Estadística comenzó su andadura sin la confluencia de la Informática. Pero actualmente se ha visto apoyada por ésta, de tal forma que, hoy en día, podemos hablar de la Estadística informatizada o de programas informáticos de Estadística, aunque debe quedar clara la necesidad imperiosa de un conocimiento importante de Estadística para poder manejar e interpretar los resultados de estos programas.

2. LA ESTADÍSTICA Y LOS MÉTODOS MULTIVARIANTES. BREVE RECORRIDO HISTÓRICO

Parece ser que la primera vez que se hace referencia explícita a la Estadística es en Alemania, en 1748, por Achenwall, aunque su anti-

güedad es muy superior. Tácito cuenta que Augusto hizo proceder a una gran encuesta sobre las riquezas del Imperio; enumeró todos los soldados, los navíos, toda suerte de recursos y las rentas públicas.

A lo largo de la Edad Media, y hasta principios del siglo xvii, la Estadística fue solamente descriptiva. Más tarde coexistieron dos escuelas: la descriptiva alemana, en la que estaba a la cabeza Achenwall, considerado por sus compatriotas padre de la Estadística, y la escuela de matemáticos políticos, que ensayó métodos de previsión y leyes sobre regularidad de ciertos fenómenos sociales.

En el siglo xviii, el francés Desparcieux y el sueco Wargentin construyeron las primeras tablas de mortalidad, pieza fundamental en las investigaciones realizadas por las compañías de seguros. En 1801 se efectuó en Francia el primer censo de población.

Jacques Bernoulli, y sobre todo Laplace, buscaron la utilización de los resultados del cálculo de probabilidades cambiando el método descriptivo por el método matemático. El primero formuló la «ley de los grandes números»; el segundo, en su teoría analítica de las probabilidades, puso en evidencia las ventajas que podía aportar el cálculo de las probabilidades en el estudio de los fenómenos naturales de causas complejas.

Durante el siglo xx la Estadística ha experimentado tal avance que los fenómenos aleatorios pueden decirse que son considerados no sólo parcela de la ciencia, sino patrimonio de la cultura, y un modo de pensar cotidiano. Dos de los artífices de este progreso fueron Sir Ronald Fisher y Abraham Wald.

Sir Ronald Fisher (1890-1962) fue uno de los científicos más eminentes en los campos de la Genética y la Estadística. Alrededor de los años veinte se ocupó del diseño de experimentos en agricultura, y de su trabajo pionero nació la teoría moderna sobre el diseño de experimentos. También, fue el introductor del análisis de la varianza.

Abraham Wald (1902-1950) fue un matemático que comenzó muy tarde sus trabajos en Estadística, aplicando los mismos a necesidades derivadas de la Segunda Guerra Mundial. Durante sus doce años de trabajo en el campo de la Estadística creó el análisis secuencial y la teoría de la decisión.

De la lista de ilustres fundadores de la Estadística actual no pueden olvidarse nombres como: Benzecri, con su análisis de datos, Borel, Cochran y Neyman, con su influencia en la inferencia estadística, Kol-

mogoroff y sus aportaciones al tratamiento de los procesos estocásticos, Savage y su teoría de juegos...

Respecto al análisis multivariante, sus orígenes se remontan a las primeras generalizaciones de la correlación y regresión por Francis Galton (1900) y Karl Pearson (1901). Este último esboza en su obra las primeras ideas del análisis de componentes principales. Otro de los trabajos pioneros se debe a Spearman (1904).

Durante el primer tercio del siglo xx se producen importantes avances en la Estadística, que posteriormente, posibilitarán el desarrollo del análisis multivariante. Cabe destacar las contribuciones de Fisher en el análisis de la varianza; Neyman en la razón de verosimilitud, que tanta influencia tendrá en la inferencia estadística y en las distribuciones multivariantes, y Pearson, en ji-cuadrado, correlación y proximidades.

En los años treinta del siglo xx, las aportaciones de Hotelling (1931, 1933), Wilks (1932, 1935), Fisher (1935, 1936), Mahalanobis (1936) y Bartlett (1939) consolidan el análisis multivariante.

Los comienzos del análisis multivariante fueron de tipo teórico, desanimando su aplicación por lo arduo de sus cálculos. Posteriormente fue Rao (1952), de la escuela hindú, quien comenzó las primeras aplicaciones prácticas.

Hasta los años setenta del siglo xx se han sucedido varias escuelas: la **inglesa** con Galton, Pearson, Fisher, Student y Snedecor; la **americana**, de Hotelling, Wilks y Bartlett, entre otros; la **hindú** con Mahalanobis, Rao, Roy, Krishnaiah, etc. En los años sesenta surge la escuela **francesa** de tratamiento de datos, con Benzecri, Lebart, Morineau y Fénélon, entre otros, y en los setenta la escuela **sueca** de Jöreskog y Sörbom, con su análisis causal.

El nacimiento de los ordenadores en los años 60 marca el despegue final del análisis multivariante. La complejidad del proceso de cálculo pasa a segundo plano y comienza a aplicarse el análisis multivariable a la Psicología, la Educación, la Biología, la Medicina, la Economía, la Geología, etc. Sin embargo, en nuestro país, la utilización de los ordenadores en la investigación empírica, salvo escasas excepciones, no se inicia hasta los años 70 y alcanza su mayoría en los 80, cuando la Informática pasa a ser de dominio público. Toda la estadística teórica de los métodos multivariantes actuales fue desarrollada antes de la aparición de los ordenadores, pero sólo cuando la informática facilitó el manejo llegó a conocerse la existencia de esas técnicas fuera del círculo de los

estadísticos teóricos (Hair, 1999, p. 3). Desde ese momento, la mayoría de investigaciones de carácter empírico suelen utilizar el análisis multivariante.

Durante los años 90 se han consolidado los métodos multivariantes en la investigación empírica, merced a los importantes logros de los paquetes estadístico-informáticos, a cuya cabeza figura el SPSS de la firma Micromouse. Estos programas, por su facilidad de uso y su acusado grado de apoyo mediante asistente estadístico, hacen «amigable» el trabajo del investigador.

Por otra parte, en ésta última década, circula un creciente empuje de la inteligencia artificial en su aplicación en los modelos expertos, la lógica borrosa y el redescubrimiento de la estadística bayesiana, que está haciéndose notar en métodos clásicos como el cluster y el discriminante, entre otros. Además los llamados métodos no lineales, del grupo de Gifi (Universidad de Leiden) facilitan la generalización de las aplicaciones de los métodos multivariantes.

3. CONCEPTO DE ANÁLISIS MULTIVARIANTE

En primer lugar debemos entrar en una cuestión terminológica. Es necesario aclarar las analogías del término análisis multivariante con otros como: análisis multivariado o análisis multivariable. Según comenta F. Azorin (1990, pp. 451-454), la primera vez que se citó, en español, un término con las connotaciones del referido, es en la obra de Barceló (1964) donde aparece como «análisis multivariante». Posteriormente todos los textos que aparecen siguen esta denominación. Hope (1972) y Cuadras (1991), e incluso la Real Academia de Ciencias Exactas, Físicas y Naturales, en su Vocabulario Científico y Técnico (1983), continúan en la misma línea. La justificación que se encuentra a esta denominación es el paralelismo con su expresión inglesa donde existe diferencia entre «variable» (con connotaciones deterministas) y «variate» (para variable aleatoria); por extensión a más de una variable surge el término «multivariate» que españolizado será multivariante. En el ámbito de las Ciencias Sociales, una de las primeras obras en tratar esta temática es la de García Ferrando (1985), que cambia la terminología por «análisis multivariable». En Ciencias de la Educación, Bisquerra (1990) y Sierra Bravo (1994) se inclinan por «multivariable» y Escotet (1979) por «multivariado»; además, en referencias de investigaciones aparece indistintamente el término «multivariante» o «multivariado», por tanto, nosotros apos-

tamos por cualquiera de los dos, por ser los más utilizados en Educación.

Si entramos en la red de redes (Internet) para conocer qué término de los tres es más utilizado, y miramos en los buscadores de mayor uso, Altavista, Yahoo y Google, los resultados son los siguientes (ver **tabla 1.1**). Como se puede observar el término más utilizado es multivariante.

TABLA 1.1. Número de páginas web donde aparecen los términos multivariante, multivariable y multivariado, en los buscadores indicados (consulta realizada en julio de 2001)

Término/buscador	Altavista	Yahoo	Google
multivariante	1992	1050	3670
multivariable	844	681	1780
multivariado	934	891	2510

A nuestro juicio, ante el debate que pueden suscitar estas cuestiones terminológicas, no debemos entrar en el juego de la famosa fábula de Iriarte, de si son galgos o podencos y pararnos en la polémica. Por consiguiente, a continuación abordaremos el concepto del análisis multivariado /multivariable /multivariante (generalmente se utilizará el último por ser el más universal).

El término análisis multivariante, diseccionado en sus componentes, tiene connotaciones de estudio minucioso (análisis), de multiplicidad (multi) y de variedad (variante).

Para Cuadras (1991, p. 1) «el análisis multivariante es un conjunto de métodos estadísticos y matemáticos para analizar, describir e interpretar las observaciones multidimensionales, es decir, el material estadístico que proviene de la observación de más de una variable».

Kendall (1975) lo define como «el conjunto de técnicas estadísticas que analizan simultáneamente más de dos variables en una muestra de observaciones».

Bisquerra (1989, p. 1) hace referencia a las técnicas multivariadas que se aplican al análisis de muchas variables y plantea como objetivos de las mismas los de resumir y sintetizar grandes conjuntos de datos.

En la definición de análisis multivariante algunos autores exigen que las variables sean aleatorias e interrelacionadas, de modo que sus

efectos no se puedan estudiar aisladamente. Es decir, se piensa que las variables son combinaciones lineales (Hair, 1999).

Recogiendo las aportaciones de distintos autores (Cuadras, Bisquerra, Kendall, Hair...) Definimos el análisis multivariante como: **el conjunto de métodos estadísticos para el tratamiento conjunto de variables con el fin de analizar, describir e interpretar datos multidimensionales.**

Destacamos dos elementos determinantes en la definición de análisis multivariante: *a)* el **carácter multidimensional de la matriz de datos** y *b)* el objetivo de **tratar ésta, conservando su estructura compleja**. Este objetivo se fundamenta en la creencia de que las variables están relacionadas entre sí, de manera que sólo el examen conjunto de las mismas puede aportar una visión completa de la realidad de forma parsimoniosa y no reduccionista.

El tratamiento conjunto está justificado porque **el todo no es igual a la combinación de las partes**, como lo demuestra el clásico ejemplo del análisis de la varianza, donde el estudio, mediante la prueba F , de la significación de la igualdad de medias de los grupos, no resulta similar a las correspondientes combinaciones de pruebas t de Student. De esta forma, **el tratamiento conjunto de las variables será el que refleje de forma fiel la realidad del problema abordado.**

4. TIPOS DE VARIABLES Y CLASIFICACIÓN DE LOS MÉTODOS MULTIVARIANTES

Cuando en la investigación educativa utilizamos el análisis multivariante, las variables que manejamos pueden clasificarse, por lo general, en independientes o predictoras, dependientes o explicadas y, cuando tienen el mismo estatus, en interdependientes.

Atendiendo a la escala de medida, se pueden presentar cuatro opciones: nominal, ordinal, de intervalo y de razón. Por el contexto de utilización de la escala de medida es interesante el agrupamiento de técnicas multivariantes; de esta manera, para facilitar su clasificación, agrupamos las escalas nominal y ordinal constituyendo la escala categórica o **no métrica**, y las escalas de intervalo y razón que integran la escala **métrica**. Las variables se denominan cualitativas, en el primer caso, y cuantitativas, en el segundo. Los valores que pueden tomar las variables son: **continuo**, cuando su recorrido así lo es, y **discreto**, si sus valores forman parte de un conjunto finito, que tiene como caso

singular el de las variables dicotómicas, cuyo conjunto está formado por dos valores. Además, si queremos convertir las variables cualitativas en cuantitativas utilizaremos las denominadas variables *dummy*, que toman el valor 1 si la característica a observar está presente y 0 si no lo está.

Algunos de los métodos estadísticos multivariantes han surgido como extensión de los univariantes que realizan una función similar; tal es el caso del análisis de la varianza como extensión de la *t* de Student, del análisis de regresión múltiple como prolongación del simple, o el análisis de segmentación (*chaid*), como generalización de la prueba de *ji-cuadrado*. Otros han surgido como variación de otros métodos multivariantes, para tratar una tipología distinta de variables; éste podría ser el caso de los modelos log-lineales y regresión logística respecto al modelo de regresión múltiple. Finalmente hay otros, como los modelos de clasificación, con cierta «autonomía».

Existen múltiples clasificaciones de los métodos multivariantes. Del análisis de las propuestas de Sánchez Carrión (1984), Tejedor (1988), Bisquerra (1989), Cuadras (1991) o Ferrán (1997), se puede afirmar, sin caer en la exageración, que cada autor ha presentado su clasificación.

Los que utilizan como criterio clasificador la dependencia de variables, clasifican los métodos en:

- Métodos de **dependencia**, donde estarían los métodos de: regresión, discriminante, multivariante de la varianza y correlación canónica.
- Métodos de **interdependencia**, con los siguientes métodos: factorial, cluster, multidimensional de escala, análisis causal.

Con un criterio similar de clasificación, Prieto (1985), citado por Bisquerra, clasifica los métodos en **predictivos**: regresión, causal, análisis múltiple de la varianza, discriminante; y **reductivos** o de interdependencia: factorial, correlación canónica, cluster, análisis de correspondencias.

Nosotros, en una labor de síntesis y teniendo como criterio clasificador el objetivo del método en la muestra o población, presentamos a continuación nuestra clasificación, que se aproxima a las presentadas por Bisquerra (1989) o Ferrán (1997):

- a) Métodos descriptivos:
 - Análisis factorial-descriptivo