MEMORIAS

I SIMPOSIO
SOBRE
METODOS ESTADISTICOS
APLICADOS A LAS CIENCIAS

Auspiciado por:

- CONICIT
- Vicerrectoría de Investigación de la Universidad de Costa Rica
- Escuela de Matemática de la Universidad de Costa Rica
- Ministerio de Relaciones Exteriores de Francia

JACQUES BADIA

JORGE POLTRONIERI

YVES SCHEKTMAN

En los últimos 15 años, la estadística matemática ha tenido un fuerte desarrollo y en particular el " Análisis de datos ".

Bajo el nombre de Análisis de datos se esconde una concep - ción simple y relativamente nueva de la estadística descriptiva la cual, apoyándose en una herramienta matemática puramente algebraica, tiene por objeto describir, reducir, clasificar, obser vaciones multidimensionales.

El desarrollo de la informática y las computadoras en los últimos 30 años han hecho posible, no solamente la explotación rápida de datos numerosos, sino también ha ayudado al desarrollo de una gran cantidad de técnicas de tratamiento de datos, tales como el análisis en componentes principales, el análisis de correspondencia, la clasificación automática, etc. Estas técnicas permiten descubrir en los fenómenos estudiados, es - tructuras directamente visibles, que no eran evidentes en los datos originales. En vista de lo anterior, surgió la necesidad de introducir estas técnicas en Costa Rica.

Como primer paso nos dimos a la tarea de organizar el 1° simposio de métodos estadísticos aplicados a las ciencias. No solo con la idea de mostrar su importancia sino también de mostrar que estos métodos se pueden utilizar en las disciplinas más diversas y se pueden atacar los problemas más complejos. En la exposición hemos tratado de dejar en claro los aportes que puede brindar el análisis de datos, utilizando problemas concretos de múltiple naturaleza: teledetección. agronomía , economía, ecología, etc. Hemos introducido, por medio de ejemplos reales, los diferentes métodos, haciendo comentarios sobre la interpretación de los resultados, teniendo en mente

la idea de presentar de la manera más clara el aspecto filosófico y dejar en un segundo plano el aspecto matemático, para la mejor comprensión y apreciación de los métodos.

No queriendo dejar los métodos en un plano puramente intuitivo, hemos presentado al principio de cada tema abordado un resumen de los resultados teóricos más importantes.

El lector podrá sin ninguna dificultad, pasar directamente a los ejemplos. Los que deseen profundizar más sobre algún tema pueden consultar los trabajos citados en la bibliografía.

———o———

El contenido científico de este documento corresponde a las exposiciones presentadas por J. Badia, Director del Laboratorio de Biometría del INRA en Tolouse, Francia y Y. Schektman, responsable cinetífico del Centro de Aplicación de la Estadística de la Universidad Paul Sabatier, Toulouse, Francia, en el Primer Simposio de Métdos Estadísticos Aplicados a las Ciencias.

Este documento ha sido redactado con la colaboración de J. Poltronieri, Profesor de la Escuela de Matemática de la Universidad de Costa Rica.

Agradecemos la colaboración brindada por el Prof. Bernardo Montero B.

Comité organizador.

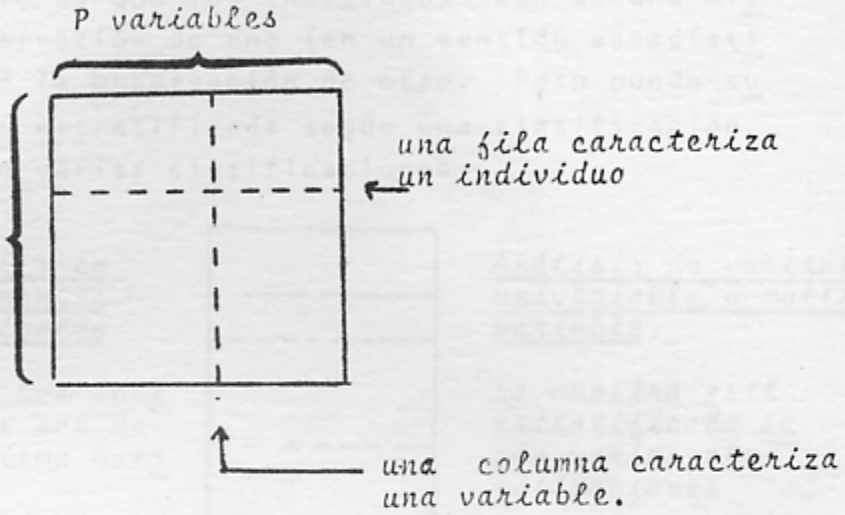
INDICE GENERAL

INTRODUCCION

GENERAL

En esta exposición de métodos estadísticos consideraremos únicamente cuadros de datos rectangulares, en los cuales

- las filas serán consideradas como los individuos de una población
- las columnas serán consideradas como las distintas caracterizaciones que se tienen sobre los individuos (variables).



Por supuesto que hay varios tipos de estos cuadros según la naturaleza de las variables. De este modo vamos a describir los cuadros más corrientes y los tipos de análisis asociados, según las preguntas planteadas. Así podemos encontrar varios criterios de diferenciación de cuadros de datos.

I - CRITERIOS DE DIFERENCIACION (4)

i-1. Según los valores de los datos

Los valores que se le pueden asociar a los individuos para una misma variable pueden ser:

- continuos ($\in \mathbb{R}$) o cuantitativos
- discretos pero ordenados ($\in \mathbb{N}$)
- cualitativos pero sin orden (color, forma, partido político)
- cualitativos binarios o indicadores. Existen solo dos alternativas: (si, no),(blanco, negro),(recto curvo), etc...

En el conjunto de las variables podemos distinguir cuatro clase de variables que llamaremos:

- explicativas (o factores o independientes)
- a explicar (o dependientes)
- concomitantes (o covariables)
- instrumentales

I-2. Según la estructura del conjunto de individuos

El caso más sencillo es que los individuos sean de una misma muestra y que la observación de uno (en un sentido estadístico) sea la repetición de la observación de otro. Pero puede suceder que la muestra sea estratificada según una clasificación, o más generalmente según varias clasificaciones.

Análisis en componentes principales

Todos los individuos son de una misma muestra.

Análisis de varianza univariable o multivariable.

La muestra está estratificada según varias clasificaciones

Es de notar que las estratificaciones pueden traducirse a un sistema binario. Cada variable binaria caracteriza un grupo de estratificación: a un individuo le corresponde un 1 si pertenece a un grupo, y un 0 en caso contrario.

Ejemplo :

Consideremos 10 individuos repartidos en 3 grupos A, B, C

V_1 representará el grupo A

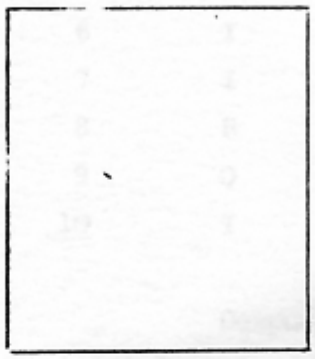
V_2 representará el grupo B

V_3 representará el grupo C

		V_1	V_2	V_3
A	1	1	0	0
	2	1	0	0
	3	1	0	0
	4	0	1	0
B	5	0	1	0
	6	0	1	0
	7	0	1	0
	8	0	1	0
C	9	0	0	1
	10	0	0	1

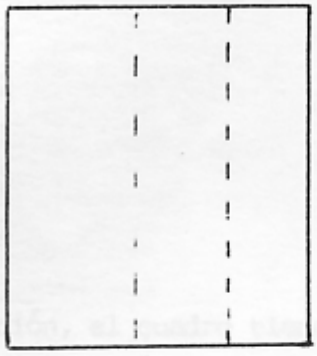
I-3. Según la estructuración del conjunto de variables

El caso más sencillo es que todas las variables sirvan únicamente para caracterizar a los individuos. Puede ocurrir sin embargo, que una variable tenga especial interés si se busca, por ejemplo, predecirla a partir de las demás. O que las variables provengan de varios orígenes, y se quiere caracterizar las relaciones que se tienen entre ellas, independencia, y en caso de dependencia, saber que tipo de dependencia existe.



Todas las variables sirven únicamente para caracterizar a los individuos

. análisis en componentes principales



Las variables no tienen el mismo interes

.análisis discriminante

.análisis de correspondencia

.análisis de regresión

Una transformación de variables de tipo cualitativo en variables de tipo binario conduce a este caso.

Por ejemplo si los trabajadores de una región se caracterizan según 3 variables :

1 - Tipo de trabajo

I : industrial

R : rural

C : comercial

O : otros casos

2 - Sexo

M : masculino

F : femenino

3 - Nivel de ingreso anual

1 : inferior a 50.000 colones

2 : entre 50.000 y 150.000 colones

3 : superior a 150.000 colones

.../...

El cuadro inicial para 10 personas podría ser :

	V_1	V_2	V_3
1	I	M	1
2	I	M	1
3	R	M	2
4	C	F	1
5	C	M	2
6	I	F	1
7	I	M	2
8	R	M	1
9	O	M	3
10	I	M	2

Después de la transformación, el cuadro tiene cada variable inicial representada con tantas variables binarias como clases tenía

	V_1				V_2		V_3		
	I	R	C	O	M	F	1	2	3
1	1	0	0	0	1	0	1	0	0
2	1	0	0	0	1	0	1	0	0
3	0	1	0	0	1	0	0	1	0
4	0	0	1	0	0	1	1	0	0
5	0	0	1	0	1	0	0	1	0
6	1	0	0	0	0	1	1	0	0
7	1	0	0	0	1	0	0	1	0
8	0	1	0	0	1	0	1	0	0
9	0	0	0	1	1	0	0	0	1
10	1	0	0	0	1	0	0	1	0

II - DISTINTOS ANALISIS ESTADISTICOS

Casi todos los métodos que consideramos son lineales, lo que explica que se puede pasar de uno a otro en el plano matemático. Pero no desde el punto de vista de interpretación de resultados.

II- 1. Análisis en componentes principales desde un punto de vista clásico.

No hay ninguna estructuración a priori de los individuos o de las variables.(1) Se trata de estructurar mediante representaciones gráfica al conjunto de los individuos. La variables deben ser continuas o por lo menos de tipo ordenado.

II-2. Análisis de varianza multivariable

Las variables son solamente un soporte para caracterizar a los individuos. Aquí los individuos llevan una estructura que puede ir de las más sencilla (un factor) a las más complicada (diseños experimentales no ortogonales). El objetivo es verificar si esa estructura tiene resonancia a priori en los valores de las variables. Las variables deben ser continuas.

II-3. Análisis de regresión

Las variables son continuas y agrupadas en dos subgrupos : las variables a explicar y las variables explicativas. Se trata de predecir unas en función de las otras. El caso es bien conocido cuando hay una sola variable a explicar.

II-4. Análisis discriminante y análisis de correspondencias

Estos tipos de análisis serán tratados pero como casos particulares del análisis en componentes principales. Algunos aspectos del análisis discriminante seran presentados en el cuadro del análisis de varianza multivariable. De la misma manera veremos que el análisis de covarianza no es diferente en su filosófia del análisis de varianza.

III - GENERALIDADES SOBRE EL ANALISIS DE DATOS

En el análisis de datos el empleo de métricas y de representaciones geométricas es fundamental para poder analizar y describir los cuadros de datos de n filas y p columnas de la forma siguiente :

.../...

(1) En el modelo clásico

$$X = \begin{bmatrix} x_{11} \cdots\cdots x_{1j} \cdots\cdots x_{1p} \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ x_{i1} \cdots\cdots x_{ij} \cdots\cdots x_{ip} \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ x_{n1} \cdots\cdots x_{nj} \cdots\cdots x_{np} \end{bmatrix}$$

Si consideramos las filas de X, el individuo i puede ser representado por sus coordenadas $(x_{i1}, \ldots, x_{ip})$ en un espacio de p dimensiones (una por variable) El conjunto de individuos constituye una "nube", denotada M de n individuos en un espacio de p dimensiones que llamaremos espacio de individuos y que notaremos E.

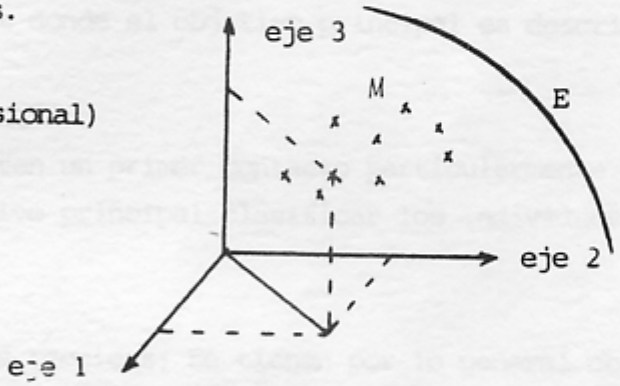
En el caso p = 3 esta nube es la nube clásica utilizada para estudiar las relaciones entre tres variables.

E : espacio de individuos (p dimensional)

M : nube de n puntos

1 punto ⟷ 1 individuo

1 eje ⟷ 1 variable



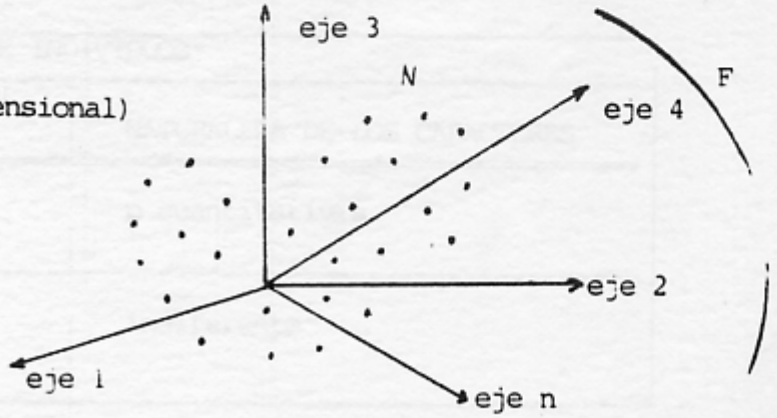
De la misma manera si consideramos las columnas de X, podemos construir una nube de p puntos en un espacio n dimensional que llamaremos espacio de las variables y que notaremos F.

F : espacio de variables (n dimensional)

N : nube de p puntos

1 punto ⟷ 1 variable

1 eje ⟷ 1 individuo



En análisis lineal, las proximidades son medidas a partir de distancias euclideanas.

... ...

Para la descripción de una nube de puntos, su forma y las posiciones de los puntos, necesitamos la definición de una distancia (i.e. una métrica). Es decir es necesario que la frase : "el individuo i está próximo al individuo i'" tenga un sentido, el cual está dado por la manera de definir dicha distancia. De igual manera se hace necesario definir la proximidad entre dos caracteres.

En resumen es la naturaleza y organización del conjunto de caracteres que permite la distinción entre las técnicas. Antes de dar una clasificación haremos algunas observaciones que permiten precisar en qué, caracteres e individuos difieren en realidad en la concepción del estadístico: dado que en el aspecto puramente matemático juegan roles simétricos.

Todas las técnicas de análisis de datos tendientes a la descripción se pueden clasificar(*) considerando :

- Por un lado : las técnicas donde el objeto principal es describir un conjunto de individuos (ningún caracter juega un rol particular) ;
- Por el otro : las técnicas donde el objetivo principal es describir las relaciones entre los caracteres.

Las primeras técnicas permiten un primer contacto particularmente eficaz con los datos. Ellas tienen por objetivo principal clasificar los individuos en grupos homogéneos.

Cuando se utilizan las otras técnicas, se tienen por lo general objetivos más precios (preveer, hacer diagnósticos, etc) sobre el problema tratado.

CLASIFICATION DE INDIVIDUOS	
T É C N I C A	NATURALEZA DE LOS CARACTERES
Análisis en componentes principales	p cuantitativas
Análisis factorial de una tabla de distancias (1)	indiferente

.../...

(1) desde un punto de vista clásico

PARTE 1

ESTADISTICA

DESCRIPTIVA

(ANALISIS LINEAL DE DATOS MULTIDIMENSIONALES)

DESCRIPCION DE RELACIONES DE CARACTERES		
TÉCNICAS	NATURALEZA DE LOS CARACTERES	
	GRUPO I (eventualmente variables a explicar	GRUPO II (eventualmente variables explicativas)
regresión multiple	1 cuantitativa	p cuantitativas
análisis de varianza	1 cuantitativa	p cualitativas
análisis de covarianza	1 cuantitativa	+ p_1 cuantitativas p_2 cualitativas
análisis factorial de correspondancias	1 cualitativa	1 cualitativa
análisis canónico[1]	q cuantitativas	p cuantitativas
análisis factorial discriminante	q cuantitativas	1 cualitativa
análisis de varianza multidimensional	q cuantitativas	p cualitativas
análisis de covarianza multidimensional	q cuantitativas	p_1 cuantitativas + p_2 cualitativas

(1) Este análisis no fue tratado directamente desde un punto de vista general en el coloquio. Pero se puede mostrar que el análisis discriminante y el análisis de correspondencia son casos particulares de éste.

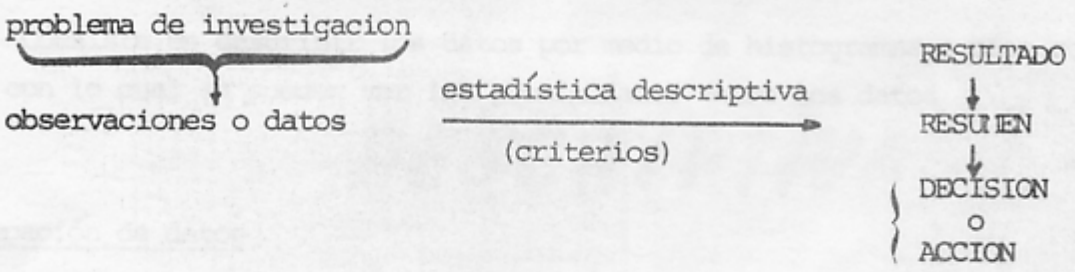
En esta sección daremos una presentación rápida de los métodos; se hará desde un punto de vista teórico y no se harán demostraciones.

Dentro de los métodos de tratamiento de datos, veremos esencialmente el análisis en componentes principales, dado que se ha demostrado que los otros métodos se pueden considerar como casos particulares de éste.

Vamos a empezar con algunas generalidades (desde un punto de vista que usualmente se admite).

I - GENERALIDADES

En el seguiente cuadro representamos un aspecto del objetivo que se persigue al investigar un problema dado :

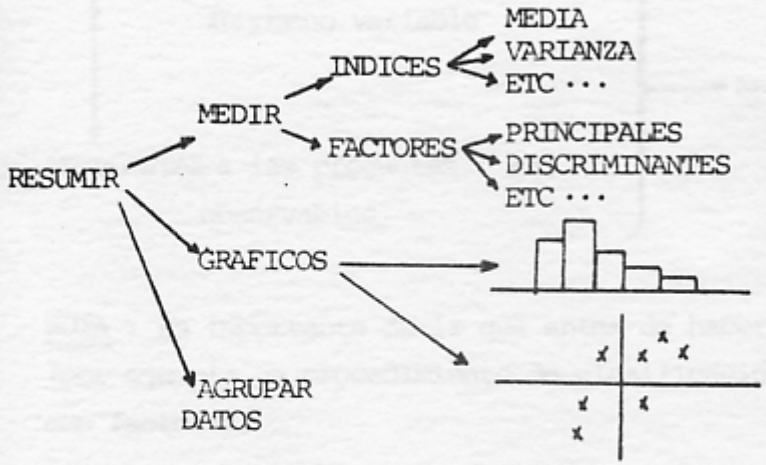


La primera fase : investigación del problema, que cubre toda une discusión sobre el fenómeno

La sugunda fase : recolección de datos u observaciones. Es en esta fase que la estadística entra en acción para sacar resumenes de los cuales se pueden hacer decisiones.

.../...

Nosotros entenderemos por resumir, tres tipos de representaciones :



i) Medir con indices

Consiste en describir los datos haciendo uso de parámetros: media, varianza, coeficiente de correlación, etc...

Medir con factores

Consiste en la descripción de los datos haciendo uso de los factores principales, discriminantes, etc...

ii) Confección de gráficos

Consiste en describir los datos por medio de histogramas,o bien con gráficos con lo cual se pueden ver las proximidades entre los datos

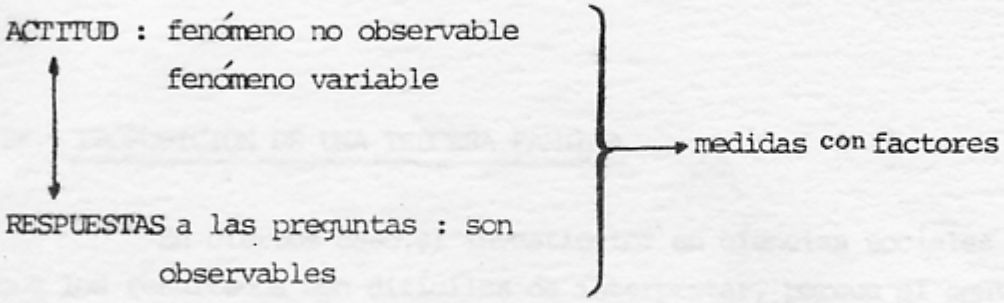
iii) Agrupación de datos

Consiste en agrupar los datos con características parecidas.
Aquí consideraremos solo medidas con factores y gráficos.

II - MEDIDAS CON FACTORES

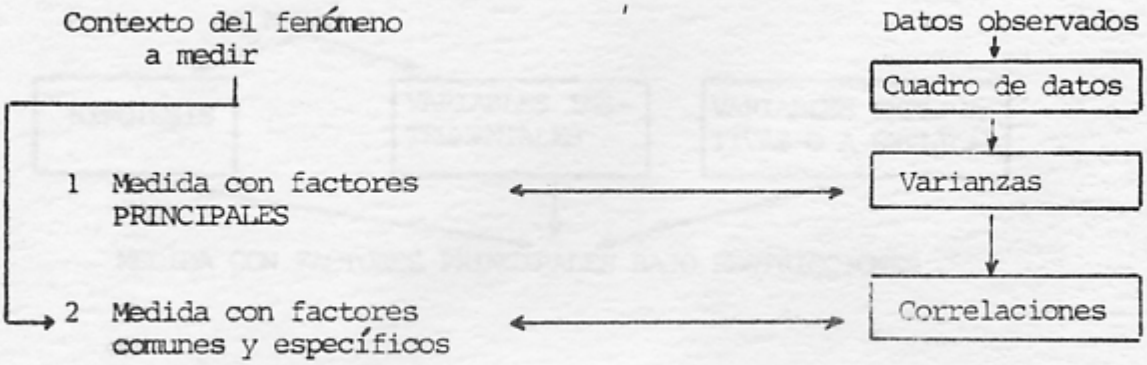
Un ejemplo : Supongamos que tratamos de medir una actitud de los individuos. Dado que la "actitud" no es un fenómeno observable, y que es un fenómeno variable, el investigator hace uso de preguntas. Claro está que las respuestas son directamente observables. En este caso se necesita medir con factores .

.../...



NOTA : Es importante decir que antes de hacer cualquier tratamiento de datos (por ejemplo un procedimiento de clasificación) es útil suministrar medidas con factores.

III - DOS GRANDES FAMILIAS DE MEDIDAS CON FACTORES



La primera familia de medidas con factores (factores principales) hace uso de las varianzas (es la noción de base), las cuales se obtienen del cuadro de datos. Las correlaciones de las variables a su vez se utilizan en la medida con factores comunes y especificos.

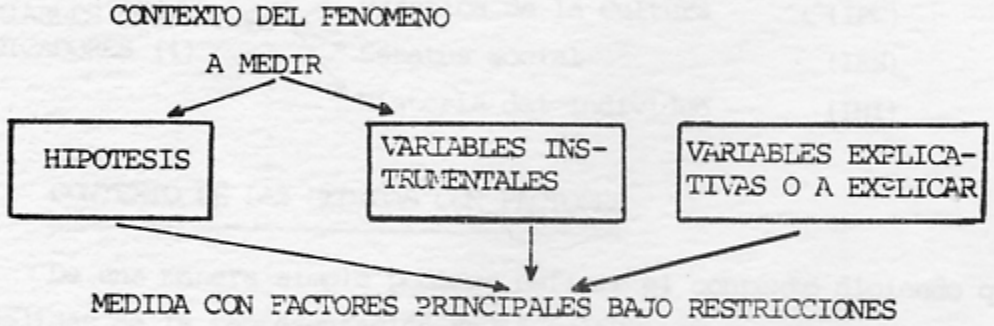
Una crítica que se hace a estas familias de medidas con factores es :
- de una manera clásica el contexto del fenómeno no es utilizado para la medida con factores principales ;
- medir con factores comunes y específicos no es el objetivo primero del modelo 2. Además la manera de cómo se utiliza el contexto del fenómeno para las medidas no es satisfactorio.

.../...

IV - PROPOSICION DE UNA TERCERA FAMILIA

En ciertos caso,el investigator en ciencias sociales se encuentra que los resultados son difíciles de interpretar, porque el contexto en el que se obtuvieron no ha sido suficientamente definido.

Dados estos resultados se ha hecho necesario estudiar el problema para volver a definir las bases EPISTEMOLOGICAS de las medidas de la tercera familia (y que contenga la primera).



PRINCIPIOS

a) Sabiendo la existencia del indeterminismo del fenómeno observado (que se traduce en ignorancia del fenómeno), se hace imperativo que el hombre tenga que contentarse con observar las variaciones del fenómeno que se estudia. Esto exige que se respete (lo mejor possible) las variaciones[1] (varianzas).

b) En general se puede decir que : "el que no sabe lo que busca, no ve lo que encuentra". Teniendo en mente esto, es necesario definir hipótesis a priori sobre las características que se quieren para las medidas con factores. De este modo tenemos que : "cualquier medida con factores toma un significado solamente cuando está de acuerdo con un sistema de hipótesis emitidas a priori".

(1) Es importante notar que elestudio de estas variaciones aportan conocimientos interesantes sobre el fenómeno estudiado en el caso que se efectue una buena escogencia de las variables y eventualmente un codificación de estas.

c) Toda medida con factores toma una significación real, solamente cuando es sacada de las relaciones existentes entre varios campos semánticos (este es el punto de vista de la escuela estructuralista).

V - UN EJEMPLO REAL (en componentes principales con restricciones).

Problema

Pretendemos estudiar la representación de la cultura en una sociedad determinada. Sabemos que dicha representación está relacionada con la práctica de la cultura, el estatus social y la historia del individuo.

LAS VARIABLES OBSERVADAS SON INDICADORES (1)

- Representación de la Cultura (IRC)
- Práctica de la Cultura (IPC)
- Estatus Social (IES)
- Historia del Individuo (IHI)

CONTEXTO DE LAS MEDIDAS CON FACTORES

De una manera simple podemos definir el contexto diciendo que queremos tener medidas de la representación de la cultura :

1 - que dependan :

de las correlaciones entre IRC y IPC
de las correlaciones entre IRC y IHI

2 - que no dependan :

de las correlaciones entre IPC y IHI
de las correlaciones internas de los IRC, IPC y IHI
del efecto de las IES

En este ejemplo hay cuatro campos semánticos. Las variables de los campos IPC, IES y IHI son instrumentales y las restricciones 1 y 2 son hipótesis a priori sobre las características que se quieren para las medidas con factores.

En el dominio de este coloquio no trataremos problema de esta complejidad ; nos limitaremos a los modelos clásicos del análisis de datos.

.../...

(1) Variables de tipo cualitativo a dos modalidades de respuesta : presencia o ausencia de una "calidad". Por ejemplo :"Lee usted novelas policíacas", las dos modalidades son SI o NO.

A - LAS BASES MATEMATICAS

I - INTRODUCCIÓN

Los métodos de análisis de datos tienen como propósito proporcionar representaciones sintéticas de una tabla de datos numéricos. Así, el análisis en factores principales permite obtener un resúmen descriptivo (bajo forma gráfica) de un conjunto de n observaciones de p dimensiones (una dimensión por variable). De este modo se dispone de una tabla rectangular de valores numéricos :

$$X_{pxn} = \begin{bmatrix} x_1^1 & \cdots & x_i^1 & \cdots & x_n^1 \\ \vdots & & \vdots & & \vdots \\ x_1^j & \cdots & x_i^j & \cdots & x_n^j \\ \vdots & & \vdots & & \vdots \\ x_1^p & & x_i^p & & x_n^p \end{bmatrix} \begin{matrix} \leftarrow x^1 \\ \\ \leftarrow x^j \\ \\ \leftarrow x^p \end{matrix}$$

$$\uparrow \quad \uparrow \quad \uparrow$$
$$x_1 \quad x_i \quad x_n$$

donde :

- la $i^{ésima}$ columna de X, x_i simboliza el individuo i
- la $j^{ésima}$ linea de X, x^j simboliza el caracter j (o variable)

Se quiere representar, en la medida de lo posible, estos datos en un espacio de dimensión pequeña, con una pérdida mínima de información.

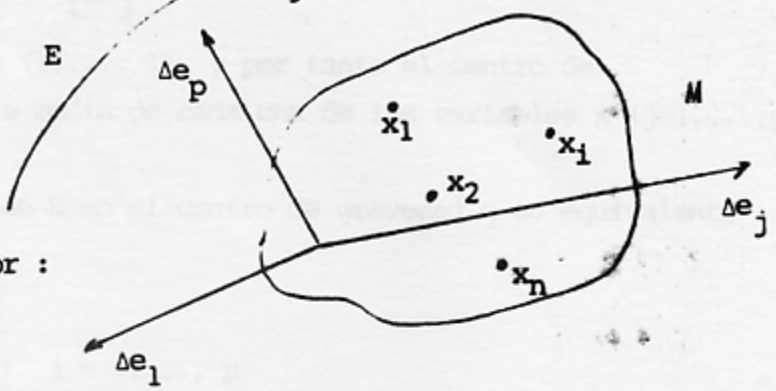
II - ESPACIO DE UNIDADES ESTADISTICAS (INDIVIDUOS - OBSERVACIONES)

El vector $x_i = \begin{bmatrix} x_i^1 \\ \vdots \\ x_i^p \end{bmatrix}$ simboliza en $E = R^p$ el individuo i, sobre el cual se han medido p variables o caracteres : E es el espacio de individuos. Estos se representan por una nube M de n puntos, si es que se tienen n individuos.

.../...

Si en (E,M) se escoge la base canónica $\{ e_j \mid j=1,\ldots,p \}$ el vector x_i se representa por :

$$x_i = \sum_{j=1}^{p} x_i^j \; e_j$$



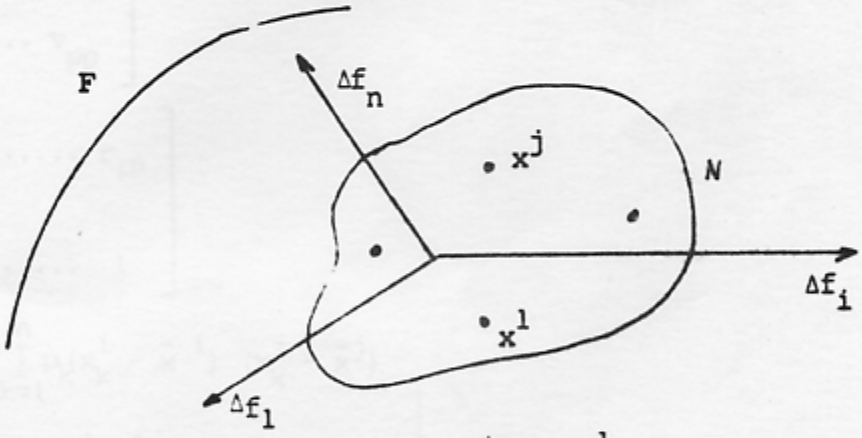
y la matriz X se puede denotar por :

$$X = (x_1, \ldots, x_n)$$

III - ESPACIO DE VARIABLES (CARACTERES)

El vector $x^j = \begin{bmatrix} x_1^j \\ \vdots \\ x_n^j \end{bmatrix}$ simboliza en $F = R^n$ el caracter j, el cual se ha medido sobre n individuos ; F es el espacio de caracteres. Estas se representan por una nube N de p puntos. Si en (F,N) se escoge la base canónica $\{ f_i \mid i = 1,\ldots, n \}$ el vector x^j se representa por :

$$x^j = \sum_{i=1}^{n} x_i^j \; f_i$$



y la matriz ${}^tX$ (transpuesta de X) se puede denotar por : ${}^tX = (x^1, \ldots, x^p)$

IV - MATRICES DE VARIANZA Y DE CORRELACION

Los elementos de la nube N se proveen de los pesos $p_i > 0$ $(i=1,\ldots,n)$ tales que :

$$\sum_{i=1}^{n} p_i = 1$$

Si consideramos en F la métrica de pesos i.e. $D_p = \begin{bmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{bmatrix}$

.../...

tenemos que el centro de gravedad de la nube $\mathcal{N}$ es :

$$g = \sum_{i=1}^{n} p_i x_i = XD_p\dot{J} = \begin{bmatrix} \bar{x}^1 \\ \vdots \\ \bar{x}^p \end{bmatrix}$$

donde $\dot{J}$ es el vector tal que ${}^t\dot{J} = (1,\ldots, 1)$, y por tanto el centro de gravedad tiene como coordenadas la media de cada una de las variables $x^j (j=1,\ldots,p)$.

Si colocamos el origen de E en el centro de gravedad g es equivalente a centrar los caracteres i.e.

$$\bar{x}^j = \sum_{i=1}^{n} p_i x_i^j = 0 \quad ; \quad j = 1,\ldots, p$$

La matriz de varianza y de correlación asociadas a la muestra son las matrices simétricas siguientes :

$$V_{pxp} = \begin{bmatrix} v_{11} \cdots\cdots v_{1p} \\ \\ v_{p1} \cdots\cdots v_{pp} \end{bmatrix}$$

$$P_{rxp} = \begin{bmatrix} 1 & r_{12} \cdots\cdots r_{1p} \\ & 1 \\ \\ r_{p1} \cdots\cdots\cdots & 1 \end{bmatrix}$$

donde $v_{ij} = \text{cov}\,(x^i, x^j) = \sum_{k=1}^{n} p_k (x_k^i - \bar{x}^i)(x_k^j - \bar{x}^j)$

$$r_{ij} = \text{corr}(x^i, x^j) = \frac{\text{cov}(x^i, x^j)}{\sqrt{\text{Var}(x^i).\text{Var}(x^j)}}$$

Notemos que los caracteres están centrados , i.e. :

$$V = X D_p {}^tX$$

$$R = D_{1/\sigma} V D_{1/\sigma}$$

con $D_{1/\sigma}$ la matriz diagonal de inversos de las desviaciones estandar. De ahora en adelante la tabla X de datos se considerará centrada, es decir que las variables x^j están centradas.

.../...

V - LAS APLICACIONES FUNDAMENTALES

Hemos considerado :

- En E la base canónica $\{e_1, \ldots, e_p\}$
- En F la base canónica $\{f_1, \ldots, f_n\}$

Así pues consideraremos :

- En E^* la base dual $\{e_1^*, \ldots, e_p^*\}$ de $\{e_1, \ldots, e_p\}$
- En F^* la base dual $\{f_1^*, \ldots, f_n^*\}$ de $\{f_1, \ldots, f_n\}$

de modo que :

$$e_i^* (e_j) = \delta_{ij}$$

$$f_i^* (f_j) = \delta_{ij}$$

i.e. $$e_j^* (x_i) = x_i^j$$

$$f_i^* (x^j) = x_i^j$$

Consideraremos las aplicaciones :

$$X : F^* \longrightarrow E \qquad y \qquad {}^tX : E^* \longrightarrow F$$

$$f_i^* \longrightarrow x_i \qquad\qquad e_j \longrightarrow x^j$$

Así definidas las aplicaciones X y ${}^tX$ tienen como matrices asociadas las tablas X y ${}^tX$ respectivamente.

VI - METRICAS EN E y F

Para describir los individuos x_i, es necesario "precisar" la noción de proximidad entre individuos.

Igualmente, para describir los caracteres x^j es necesario definir la proximidad entre dos caracteres.

.../...

En análisis lineal las proximidades se miden con distancias euclideanas, por lo que :

- para medir proximidad entre individuos, se provee al espacio de individuos E de una métrica euclideana M,
- para medir la proximidad entre caracteres y juzgar de su colinealidad, se provee el espacio de caracteres F de la métrica N .

Recordemos que una métrica euclideana Q sobre un espacio vectorial E de dimensión k es tal que :

Q es una aplicación de E x E en R^+ : $(x,y) \longrightarrow {}^t x Q y$

Notaremos $Q(x,y) = {}^t x Q y$

(1) Q es simétrica : $Q(x,y) = Q(y,x)$

(2) Q es bilineal $Q(\lambda x + \delta y , z) = \lambda Q(x,z) + \delta Q(y,z)$

$Q(x, \alpha y + \beta z) = \alpha Q(x,y) + \beta Q(x,z)$

(3) Q es definida $Q(x,x) = 0 \Longleftrightarrow x = 0$

(4) Q es positiva $x \neq 0 \Longleftrightarrow Q(x,x) > 0$

y define :

. un producto escalar : $Q(x,y) = {}^t x Q y$ es el producto escalar de x e y

. una norma euclídeana: $||x||_Q = \sqrt{Q(x,x)}$ es la Q norma del vector x

. una distancia euclideana : $d(x,y) = ||x-y||_Q$ es la Q distancia entre x e y

. una ortoganalidad : $Q(x,y) = {}^t x Q y = 0 \Leftrightarrow$ x es Q ortogonal a $y \Longleftrightarrow x \overset{\perp}{Q} y$.

.../...

Notemos que Q se utiliza como aplicación y como matriz de tal forma que :

$$Q = (q_{ij}), \text{ donde } q_{ij} = Q(e_i, e_j)$$

es la matriz de Q en la base $\{e_1, \ldots, e_p\}$.

Además Q permite asociar a todo vector $x \in E$ una forma lineal Q_x definida así :

$$Q_x : E \longrightarrow R$$
$$y \longrightarrow Q(x,y)$$

por lo que $Q_x \in E^*$

Si denotamos φ la aplicación tal que :

$$E \xrightarrow{\varphi} E^*$$
$$x \longrightarrow Q_x$$

y si consideramos en E^* la base dual $\{e_1^*, \ldots, e_p^*\}$

la matriz asociada a φ es Q, i.e. :

$$E \xrightarrow{\varphi} E^*$$
$$x \longrightarrow Q_x = {}^t x\, Q$$

de modo que $Q_x(y) = {}^t x\, Qy$

VII - FORMAS BILINEALES INDUCIDAS

VII-1. Sobre F^* por medio de M vía la aplicación X

Nos interesamos en definir sobre F^* una forma bilineal W tal que

$$||f_i^* - f_{i'}^*||_W = ||X(f_i^*) - X(f_{i'}^*)||_M = ||x_i - x_{i'}||_M$$

.../...

Así tenemos en virtud de la forma bilineal M :

$$M(x_i, x_{i'}) = {}^t x_i \, M \, x_{i'} = {}^t f_i^* \; {}^t X \, M X \, f_{i'}^* = {}^t f_i^* \; W \; f_i^*$$

i.e. $W = {}^t X \, M \, X$

VII-2. Sobre E^* por medio de N vía la aplicacion ${}^t X$

Igualmente nuestro interés es definir sobre E^* una forma bilineal V tal que :

$$||e_j^* - e_{j'}^*||_V = ||{}^t X (e_j^*) - {}^t X(e_{j'}^*)||_N = ||x^j - x^{j'}||_N$$

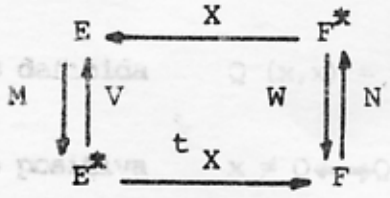
Así tenemos en virtud de la forma bilineal N :

$$N(x^j, x^{j'}) = {}^t x^j \, N \, x^{j'} = {}^t e_{j'} \, X \, N \, {}^t X \, e_j = {}^t e_{j'} \, V \, e_j$$

i.e. $V = X \, N \, {}^t X$

VII-3. El esquema de dualidad

El esquema siguiente describe las relaciones entre los diferentes espacios considerados :



(esquema de dualidad)

de modo que :

$$- W = {}^t X \, M \, X$$
$$- V = X \, N \, {}^t X$$

VIII - METRICA DE PESOS - FORMA CUADRATICA DE INERCIA

Los caracteres x^j se consideran centrados, i.e. el centro de gravedad de M coincide con el origen de E, $\vec{g} = \vec{0}$.

.../...

Si en F se considera la métrica de pesos $N = D_p$ se deduce que :

- $D_p\ (f_i, f_k) = 0 \qquad i \neq k$
- $D_p\ (f_i) = ||f_i||^2_{D_p} = p_i$
- $D_p\ (x^i, x^j) = \text{cov}\ (x^i, x^j)$
- $D_p\ (x^j) = ||x^j||^2_{D_p} = \text{Var}\ (x^j)$

Como $V = X\, D_p\, {}^tX$ se tiene que :

$$V(e_i^*, e_j^*) = D_p\ (x^i, x^j) = \text{cov}\ (x^i, x^j)$$
$$V(e_j^*) = D_p\ (x^j) = \text{Var}\ (x^j)$$

o sea que la matriz de V en la base $(e_1^*, \ldots, e_p^*)$ es la matriz de varianza-covarianza de los caracteres x^j.

———————— + ————————

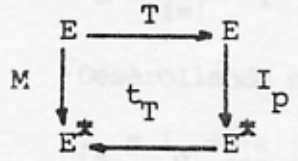
.../...

B - ANALISIS EN COMPONENTES (FACTORES) PRINCIPALES

I - INTRODUCCION

Si el ojo pudiera ver en p dimensiones (p > 3) la clasificación de individuos no presentaría ningúnproblema y la distancia entre individuos sería la longitud del segmento que los une. Así considerando una base M - ortonormada es necesario adoptar una representación cartesiana de E, procediendo de la siguiente manera :

Consideramos una aplicación lineal T tal que $M = {}^tT\, I_p\, T$ i.e.



$$z_i = Tx_i \quad (i=1,\ldots,n)$$

y la nube $T(M) = \{z_i \mid i=1,\ldots n\}$ sería la representación de M en un espacio euclídeo con la métrica clásica i.e.

$$||x_i - x_j||_M = ||z_i - z_j||_{I_p} \qquad \forall\ i,j$$

Pero el ojo no puede ver en un espacio a p dimensiones (p > 3). Así si queremos ver las proximidades, se busca una representación euclídeana más simple en un subespacio $E_1 \subset E$.

Para la representación euclidea mas simple de E utilizaremos criterios de "proximidad". Nos preguntaremos :

1 - ¿ cuál es el punto de E más proximo a la nube M ? ¿ la nube se concentra en ese punto ?

2 - ¿ cuál es la recta principal mas proxima de M ? ¿ la nube se sitúa en esta recta ?

3 - ¿ cuál es el plano principal mas proximo de M ? ¿ la nube se sitúa en este plano ?

etc...

.../...

Para responder a estas preguntas es necesario definir un índice que mida la proximidad :

- de la nube M y de un punto de E, y más generalmente
- de la nube M y de su subesespacio de E.

II - INERCIA

Inercia en un punto :

La "proximidad" (inercia) de un punto"a" a la nube M la definimos con el índice siguiente llamada la inercia de M con respecto el punto a

$$I_a = \sum_{i=1}^{n} p_i \, ||x_i - a||_M^2$$

Desarollando esta expresión tenemos que :

$$I_a = I_g + ||a||_M^2 \qquad \text{(Teorema de Huygens)}$$

donde $\vec{g} = \vec{0}$ es el centro de gravedad de la nube M . Así tenemos que I_a es mínimo cuando"a"coincide con el centro de gravedad.

Otra expresión de I_g que se puede obtener utilizando la métrica M es :

$$I_g = \text{traza } (VM)$$

Inercia con respecto a un subespacio vectorial (s.e.v)

Sea H un s.e.v. de E y sea $H^{\perp}$ el s.e.v. suplementario M ortogonal a H i.e.

$$E = H \oplus H^{\perp}$$

Sabemos que :

$$x_i = \alpha_i + \beta_i \quad \text{con } \alpha_i \in H \; ; \; \beta_i \in H^{\perp}$$

La cantidad :

$$||x_i - \alpha_i||_M = ||\beta_i||_M$$

mide la proximidad del punto x_i al s.e.v H.

.../...

Para medir la proximidad de la nube M y del s.e.v H se utiliza el indice :

$$I_H = \sum_{i=1}^{n} p_i \, ||x_i - \alpha_i||_M^2 = \sum_{i=1}^{n} p_i ||\beta_i||_M^2$$

por lo que :

$$I_{H^\perp} = \sum_{i=1}^{n} p_i ||\alpha_i||_M^2$$

Nota :

- $M \subset H \Longleftrightarrow I_H = 0$
- $I_g = I_H + I_{H^\perp}$

III - DETERMINACION DEL PRIMER EJE PRINCIPAL -PROPIEDADES - DEFINICIONES

Para la determinación del primer eje principal, se pretende encontrar un vector u de norma 1 tal que el s.e.v Δ_u (recta de vector director u) sea de inercia mínima, es decir, I_{Δ_u} sea mínima.

Dado que $I_g = I_{\Delta_u} + I_{\Delta_u^\perp}$ minimizar I_{Δ_u} es equivalente a maximizar $I_{\Delta_u^\perp}$ con la condición de que $||u||_M^2 = 1$.

Una expresión para $I_{\Delta_u^\perp}$, cuando u es M normado es :

$$I_{\Delta_u^\perp} = M \, V \, M \, (u,u)$$

por lo que tenemos que maximizar la forma cuadrática M V M (u,u) cuando $||u||_M = 1$.

El sistema se escribe, utilizando multiplicadores de Lagrange, así :
maximizar (1) $L = {}^t u \, M \, V \, M u - \lambda ({}^t u \, M u - 1)$

Derivando obtenemos que :

$$2 \, V M u - 2\lambda u = 0$$

por lo que u es un vector proprio de VM, es decir $VMu = \lambda u$

Sustituyendo en (1) obtenemos que :

$$L = \lambda \, {}^t u \, M \, u - \lambda \, {}^t u \, M \, u + \lambda = 0$$

por lo que L es máximo cuando λ es el valor proprio más grande de VM.

En resumen tenemos que : el primer eje principal es el eje Δ_{u_1} engendrado por u_1, vector proprio M normado de VM asociado al más grande valor proprio λ_1.

Al primer eje principal en E está asociado :

- en E^* la primera forma lineal principal : $v_1 = Mu_1$
- en F la primera componente (o factor) principal :

$$c^1 = {}^t X\, v_1 = {}^t X\, M\, u_1$$

Así se ve que los valores absolutos de las coordenadas del vector c^1, en la base canónica de F, son las normas de la proyección M ortogonal de los vectores $\{x_1, \ldots, x_n\}$ sobre Δu_1.

Tenemos las propriedades siguientes :

$$\cdot\ I_{\Delta u_1} = {}^t u_1\, MVMu_1 = {}^t v_1\, V\, v_1 = {}^t(c^1)\, D_p(c^1) = ||v_1||^2_V = ||c^1||^2_{D_p} = \lambda_1$$

$$\cdot\ WD_p\,(c^1) = \lambda_1 c^1 \qquad \text{donde} \quad W = {}^t X\, M\, X$$

IV - DETERMINACION DE LOS PLANOS PRINCIPALES - PROPIEDADES- DEFINICIONES

En la determinación del plano principal H debemos encontrar el[1] eje Δu engendrado por el vector u que maximiza :

$$I_{\Delta u} = M\, V\, M\,(u,u)$$

bajo las restricciones :
$$\begin{cases} \|u\|^2_M = M(u,u) = 1 \\ M(u,u_1) = 0 \end{cases}$$

El sistema se escribe :

(2) $$L = {}^t u\, M\, V\, Mu - \lambda({}^t u\, M\, u - 1) - \beta(u\, {}^t M\, u_1)$$

De este modo obtenemos :

$$\begin{cases} V\, M\, u - \lambda\, M\, u - \beta M\, u_1 = 0 \\ {}^t u\, M\, u = 1 \\ {}^t u_1\, M\, u = 0 \end{cases}$$

.../...

(1) Se supone que no hay valor propio multiple. En el caso general puede monstrar que un plano principal H contiene necesariamente un eje principal Δ_{u_1}

Se deduce que $\beta = 0$ y que $V M u = \lambda u$ con la restricción ${}^t u_1 M u = 0$ por lo tanto u es el vector proprio u_2 de VM asociado al segundo valor proprio más grande λ_2.

i.e. $V M u_2 = \lambda_2 u_2$

${}^t u_2 M u_1 = 0$

El plano principal es el s.e.v. de E : $H = \Delta u_1 \oplus \Delta u_2$

Al segundo eje principal esta asociado u :

- en E^* la segunda forma lineal principal : $v_2 = Mu_2$
- en F la segunda componente (o factor) principal :

$$c^2 = {}^t X v_2 = {}^t X M u_2$$

Se puede monstrar :

. $I_{\Delta u_2} = {}^t u_2 M V M u_2 = {}^t v_2 V v_2 = {}^t(c^2) D_p c^2 = ||v_2||^2_V = ||c^2||^2_{D_p} = \lambda_2$

. $W D_p (c^2) = \lambda_2 c^2$

. $I_{H^*} = \lambda_1 + \lambda_2$ = inercia de la proyección [illegible] de la nube M sobre el subespacio H con respecto su centro de gravedad

Procediendo de la misma manera : el s.e.v principal H de dimensión k de inercia mínima, es el s.e.v engendrado por los k primeros ejes principales : Δu_1 , Δu_2, ..., Δu_k ; donde Δu_i es el eje engendrado por el vector proprio M normado u_i de VM asociado al i^{avo} valor proprio mas grande λ_i, i.e.

$V M u_i = \lambda_i u_i$ $\qquad i = 1, \ldots, k$

${}^t u_i M u_j = 0$ $\qquad i \neq j$

.../...

El s.e.v. principal $H = \Delta_{u_1} \oplus \Delta_{u_2} \oplus \ldots \oplus \Delta_{u_k}$.

A los vectores u_i corresponden :

- en E^* las formas lineales principales : $v_i = M\, u_i$ $\quad i=1,2,\ldots,k$
- en F las componentes (o factores) principales :

$$c^i = {}^tX\, v_i = {}^tX\, M\, u_i \qquad i=1,2,\ldots,k$$

Se puede mostrar,

- $I_{\Delta_{u_i}} = {}^tu_i\, MVM\, u_i = {}^tv_i\, V\, v_i = {}^tc_i\, D_p\, c_i = \|v_i\|_V = |c^i|_{D_p} = \lambda_i$
- $I_{H^\perp} = \sum_{i=1}^{k} \lambda_i$ = inercia de la proyección M-ortogonale de la nube M sobre H, con respecto su centro de gravedad.
- $W\, D_p\, c^i = \lambda_i\, c^i \qquad i=1,2,\ldots,k$
- $M \not\subset H \iff I_H > 0 \iff \dfrac{\sum_{i=1}^{k} \lambda_i}{tr(VM)} < 1 \qquad M \subset H \iff \dfrac{\sum_{i=1}^{k} \lambda_i}{tr(VM)} = 1$
- La cantitad $\sum_{i=1}^{k} \lambda_i \,/\, tr(VM)$ es el porcentage de <u>inercia explicado</u> por el subespacio H, y mide la "cantidad de information" conservado en la proyección M ortogonale de la nube M sobre el subespacio H.

V - <u>PROPRIEDADES DE LAS COMPONENTES (FACTORES) PRINCIPALES</u>

Las componentes principales c^i son centradas y de varianza λ_i :

- $\bar{c}^i = D_p(j, c^i) = \sum_{j=1}^{n} p_j\, c^i_j = 0$
- $Var\,(c^i) = \|c^i\|^2_{D_p} = \lambda_i$
- $WD_p(c^j) = \lambda_j\, c^j$

.../...

. $\sum_{i=1}^{p} \text{Var}\,(c^i) = \text{tr}(VM) = I_g$

. $\rho\,(c^i, c^j) = 0 \qquad i \neq j$ donde ρ es el coeficiente de correlación de c^i y c^j.

Nota : es interesante de notar que

. $\text{Var}(x^j) = \sum_{i=1}^{p} \lambda_i\, (u_i^j)^2$

. $\rho^2(x^j, c^i) = \dfrac{\lambda_i\, (u_i^j)^2}{\text{Var}\,(x^j)}$ = proporción de la varianza de x^j explicada por c^i

donde u_i^j es la j^{ieme} coordenada del i^{eme} vector axial principal u_i.

VI - DESCRIPCION DE LA NUBE DE INDIVIDUOS

La calidad de la representacion de un punto x_i sobre el plano principal se mide por la comparación de las normas :

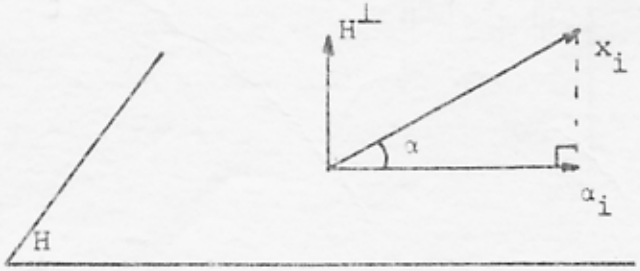
$$||x_i||^2 = \sum_{j=1}^{p} M^2(x_i, u_j) = \sum_{j=1}^{p} (c_i^j)^2$$

$$||a_i||^2 = M^2(x_i, u_1) + M^2(x_i, u_2) = (c_i^1)^2 + (c_i^2)^2$$

con la ayuda del índice :

$$\cos \alpha = \frac{||a_i||}{||x_i||}$$

donde a_i es la proyección M ortogonal de x_i sobre el plano principal H.



.../...

Esto afecta el signo del valor c_i^3 tomado por la tercera componente principal.

El examen en el plano principal de la nube de proyecciones α_i de los puntos x_i (teniendo en cuenta la calidad de la representación) permite agrupar individuos en clases de individuos semejantes.

VII - DESCRIPCION DE LOS CARACTERES

A la representación de los puntos individuos en el plano principal esta asociada la representación de los caracteres en el plano $(c^1, c^2) \subset F$.

El caracter x^i interviene más en la descripción de la nube M por su proyección en el plano principal, cuanto el valor :

$$||x^i||^2 = \sum_{j=1}^{p} \frac{D_p^2 (x^i, c^j)}{\lambda_j}$$

este próximo de $\frac{D_p^2 (x^i, c^1)}{\lambda_1} + \frac{D_p^2 (x^i, c^2)}{\lambda_2}$ = norma al cuadrado de la proyección D_p ortogonal de x^i sobre el plano (c^1, c^2).

La interpretación de las componentes principales c^1 y c^2 se hace tomando en cuenta las proximidades entre las proyecciones de los caracteres x^i y las componentes principales c^1 y c^2.

Así pues en este plano se evidencia (si las normas están bien reconstruidas) :

- proximidades entre caracteres
  entre caracteres y componentes
- ortogonalidad entre caracteres
  entre caracteres y componentes.

.../...

VIII - UN CASO PARTICULAR

Descripción de las componentes de una n serie cronológica por el análisis en componentes principales

Sobre una población de n individuos o unidades estadísticas se mide una variable z en p instantes $\tau_1, \ldots, \tau_p$. A la i^{ava} unidad estadística corresponde la observación de la i^{ava} componente $\{z_i(\tau_1), z_i(\tau_2) \ldots, z_i(\tau_p)\}$ de una serie cronológica.

Estas componentes de series cronológicas constituyen las líneas de la tabla :

$$ {}^tZ = \begin{bmatrix} z_1(\tau_1) & \ldots & z_1(\tau_p) \\ \vdots & & \\ z_n(\tau_1) & \ldots & z_n(\tau_p) \end{bmatrix} $$

Las columnas $\{z(\tau_j) \; ; \; j=1,\ldots,p\}$ de la tabla ${}^tZ$ constituyen las variables de la tabla clásica en análisis de datos.

Si $u_1, \ldots, u_k$ ($k \leq p$) constituyen los k primeros vectores axiales factoriales principales y $C^1, \ldots, C^k$ las k primeras componentes principales correspondientes se tiene :

$$ z_i(\tau_j) = \bar{z}(\tau_j) + \sum_{\ell=1}^{k} C_i^{\ell} \, u_{\ell}(\tau_j) + \text{error}_i(\tau_j) \qquad j = 1, \ldots, p $$

donde

- $\bar{z}(\tau_j)$ la media de la variable $z(\tau_j)$ sobre las n unidades estadísticas,
- C_i^{ℓ} el valor de la 1^{ava} componente principal (o factor principal) de la i^{ava} unidad estadística,
- $u_{\ell}(\tau_j)$ la j^{ava} coordenada del 1^{avo} vector axial factorial principal.

.../...

En el contexto particular de utilización del A.C.P. es útil representar por curvas las matrices columnas siguientes :

$$\begin{bmatrix} z_i(\tau_1) \\ \vdots \\ z_i(\tau_p) \end{bmatrix} \quad \begin{bmatrix} \bar{z}(\tau_1) \\ \vdots \\ \bar{z}(\tau_p) \end{bmatrix} \quad \begin{bmatrix} u_1(\tau_1) \\ \vdots \\ u_1(\tau_p) \end{bmatrix} \quad \begin{bmatrix} u_k(\tau_1) \\ \vdots \\ u_k(\tau_p) \end{bmatrix} \quad \begin{bmatrix} \text{error}_i(\tau_1) \\ \vdots \\ \text{error}_i(\tau_p) \end{bmatrix}$$

1 2 3 4

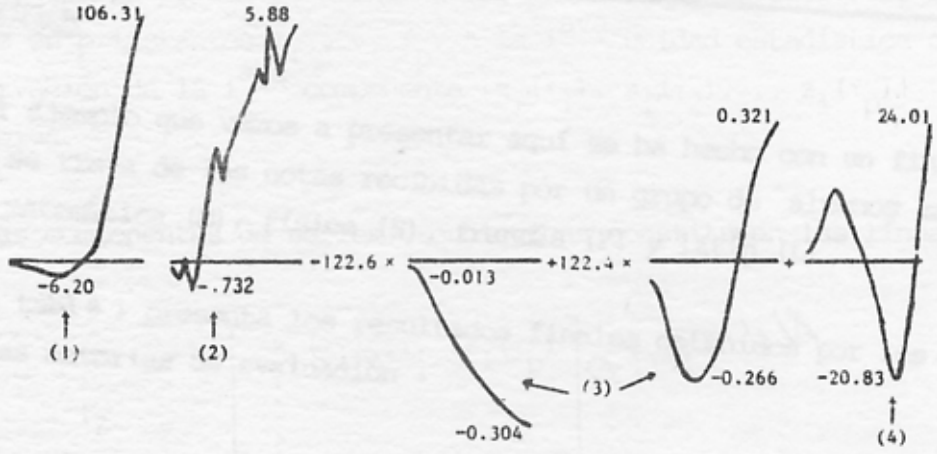
Aplicado en una n serie cronológica el A.C.P. nos lleva a descomponer, para cada unidad estadística, la curva de la serie cronológica correspondiente (1) en la suma :

- de la curva media de las series cronológicas (2),
- de las curvas (3), o contornos principales, ponderados por el valor de la componente principal correspondiente de la unidad estadística considerada,
- de una curva de error (4).

Damos como ilustración la descomposición obtenida por una empresa en el segundo A.C.P. del segundo ejemplo en el paragrafo C. para k = 2

(1)		(2)		(3)		(4)
-1.57		-0.35		-0.013		-0.036		1.5926
-2.99		-0.48		-0.021		-0.058		2.0146
-3.33		0.15		-0.036		-0.099		4.224
-3.89		-0.73		-0.050		-0.143		8.2132
-4.10		-0.55		-0.073		-0.198		11.7354
-4.89		0.29		-0.091		-0.229		11.693
-4.76		0.83		-0.107		-0.246		11.4022
-6.20		1.96		-0.129		-0.257		7.4814
-5.16		2.88		-0.148		-0.266		6.3736
-5.65		2.03		-0.168		-0.265		4.1592
-3.64		2.86		-0.182		-0.262		3.2556
-1.40	=	3.16	+(-122.6)	-0.197	+122.4	-0.246	+	1.3982
0.33		3.31		-0.207		-0.216		-1.9198
1.27		3.73		-0.214		-0.191		-5.318
4.41		4.09		-0.227		-0.148		-9.395
5.53		4.52		-0.236		-0.095		-16.2956
14.15		4.70		-0.251		-0.004		-20.833
29.42		4.56		-0.260		0.065		-14.972
40.78		5.88		-0.274		0.128		-14.3596
57.04		5.12		-0.281		0.178		-4.3178
68.98		4.47		-0.287		0.221		2.2734
81.94		5.26		-0.293		0.262		8.6894
92.89		5.66		-0.301		0.298		13.8522
106.31		5.74		-0.304		0.321		24.0092

En la figura que sigue hemos representado las curvas asociadas a las matrices columnas de la tabla anterior. Las curvas no tienen la misma escala. Los valores mínimos y máximos se indican en cada curva.



Cuando la representación de la unidad estadística sobre el plano principal es buena, se constata que, entre más se aleja del origen del gráfico (centro de gravedad de la nube de unidades estadísticas) mas los coeficientes de CONTORNOS (3) son grandes, por lo tanto explica mejor la forma de la curva (1).

En el ejemplo las coordenadas (-122.6, 122.4) sobre el primer plano principal indican (dado que son grandes y muy próximas) que la forma de la curva se asemeja a la curva asociada a $u_1 + u_2$.

C - EJEMPLOS DE ANALISIS EN FACTORES (COMPONENTES) PRINCIPALES

PRIMER EJEMPLO DE ANALISIS EN FACTORES PRINCIPALES

NOTAS ESCOLARES - EJEMPLO PEDAGOGICO

I - INTRODUCTION

El ejemplo que vamos a presentar aquí se ha hecho con un fin pedogógico. Se trata de las notas recibidas por un grupo de alumnos en las materias de matemática (M), física (S), francés (F) y latín (L).

La tabla 1 presenta los resultados finales obtenidos por los estudiantes en las materias de evaluación :

	M	S	F	L
JUA	6.0	6.0	5.0	5.5
ALA	8.0	8.0	8.0	8.0
ANA	6.0	7.0	11.0	9.5
DID	14.0	14.0	12.0	12.5
AND	11.0	10.0	5.5	7.0
MON	14.5	14.5	15.5	15.0
PED	5.5	7.0	14.0	11.5
BRI	13.0	12.5	8.5	9.5
EVE	9.0	9.5	12.5	12.0

Tabla 1
cuadro de datos)

En la tabla 1 aparecen solo las 3 primeras letras de los nombres de los estudiantes, pues solo se han asignado tres plazas en la computadora para este efecto. Esta tabla representa el cuadro de datos, y denotaremos por ${}^tX$ la matriz correspondiente (ver resultados teóricos). Las unidades estadísticas son los estudiantes : Juan, Alan, Ana, Didier, Andrés, Mónica, Pedro, Brigitte y Evelyne (n : el número de unidades estadísticas es igual a 9). Las variables son las materias que representamos por el conjunto (M,S,F,L).

.../...

II - ANALISIS DE LOS RESULTADOS A.C.P.

II-1. Escogencia de la métrica D_p :

En el estudio se le da una importancia igual a cada uno de los estudiantes. Por esta razón hemos escogido la métrica $p_i = \frac{1}{9}$ (i=1,...,9) i.e. D_p es la matriz diagonal en el espacio de variables $F(=R^9)$ es :

$$D_p = \begin{bmatrix} 1/9 & & 0 \\ & \ddots & \\ 0 & & 1/9 \end{bmatrix}$$
$$(9,9)$$

En el estudio las variables se centraron por lo que la nueva variable x'^j se escribe : $x'^j = x^j - \bar{x}^j$ donde x^j est la variable $j^{ésima}$ del cuadro inicial (tabla 1) y $\bar{x}^j$ es el vector que tiene todas sus componentes iguales a la media de las notas de la j^{esima} variable. Así el centro de gravedad de la nube de individuos está en el origen del espacio de unidades estadísticas $E(=R^4)$. El hecho de centrar la nube (M) no cambia la forma de (M) ni las proximidades entre individuos.

Así por ejemplo tenemos que la media de M es 9.67 y por tanto los valores centrados de M se representan por el vector :

(6-9.67, 8-9.67, 6-9.67, 14-9.67, 11-9.67, 14.5-9.67, 5.5-9.67, 13-9.67, 9-9.67) = (-3.67, -1.67, 3.67, 4.33, 1.33, 4.83, -4.17, 3.33, -0.67)

La matriz de varianzas y covarianzas de las variables se dan en la tabla 2 .

	M	S	F	L
M	11.40	9.92	2.66	4.82
S	9.92	8.94	4.12	5.48
F	2.66	4.12	12.10	9.29
L	4.82	5.48	9.29	7.91

Tabla 2

.../...

Esta matriz es calculada por la expresion ${}^tX' D_p X'$ donde : X' es el cuadro de observaciones con las variables centradas : x'^j ; ${}^tX'$ la matriz transpuesta de la matriz X'.

Así

$$\text{Var}(x^j) = D_p(x'^j, x'^j) = ||x'^j||^2_{D_p} = {}^t x'^j D_p x'^j$$

$$\text{cov}(x^j, x^k) = D_p(x'^j, x'^k) = \langle x'^j, x'^k \rangle_{D_p} = {}^t x'^j D_p x'^j$$

Demos un ejemplo de cálculo de varianza :

$\text{Var}(M) = {}^tM' D_p M'$ i.e.

$$[-3.67, \ldots, -0.67] \begin{bmatrix} 1/9 & & 0 \\ & \ddots & \\ 0 & & 1/9 \end{bmatrix} \begin{bmatrix} -3.67 \\ \vdots \\ -0.67 \end{bmatrix}$$

$$= \sum_i^9 p_i M_i'^2 = \frac{1}{9} (102.50) = 11.4$$

II-2. Escogencia de la métrica M

En el espacio $E(=R^4)$ escogemos como métrica M la métrica euclideana usual i.e. M=I.

De la teoria se sabe que los vectores propios(u_j) de V M = V, es decir, de la matriz de varianzas y covarianzas son los vectores axiales factoriales. En este caso tenemos que los valores propios (λ_j) de la matriz V son muy pequeños para $j \geq 3$ (tabla 3).

.../...

λ_j	1	2	3
Valor proprio	28.235	12.031	0.043
Porcentage inercia explicado	70.0	29.8	0.2
Porcentage acumulado	70.0	99.8	100.0

Tabla 3

Sabemos que la varianza total (inercia) es igual a :

$$I(M/G) = \text{traza } (V\,M) = \text{traza } (V) \overset{(1)}{=} Var(M) + Var(S) + Var(F) + Var(L)$$

$$= 40.360 = \sum_{j=1}^{4} \lambda_j$$

y que el porcentage de inercia explicada por los dos primeros ejes factoriales es del 99.8 % por lo que la nube M esta prácticamente situada en el plano engendrado por los dos primeros ejes.

II-3. Resultados análiticos en función de los factores principales centrados

Sabemos que $x^j = \bar{x}^j + \sum_{k=1}^{p} u_k^j \, c^k$, es decir que podemos reconstruir las variables a partir de los factores. Asi tenemos (2)

$$M \simeq 9.67 + .515 \times c^1 - .569 \times c^2$$

$$S \simeq 9.83 + .508 \times c^1 - .371 \times c^2$$

$$F \simeq 10.20 + .492 \times c^1 + .658 \times c^2$$

$$L \simeq 10.10 + \underset{\uparrow}{\underset{u_1}{.484}} \times c^1 + \underset{\uparrow}{\underset{u_2}{.325}} \times c^2$$

La primera columna de la derecha representa las coordenadas del centro de gravedad en el espacio inicial) y las otras las coordenadas de los vectores axiales factoriales principales.

.../...

(1) Esta igualdad tiene porque M es la métrica clásica I.

(2) Aquí los dos últimos factores explican muy poco de inercia. Con los dos primeros se puede reconstruir las variables con muy pequeños errores.

II-4. Resultados analiticos en función de los factores principales reducidos

Sabemos que la varianza de c^j es igual a λ_j i.e. expresaremos las variables x^j de la siguiente manera :

$$x^j = \bar{x}^j + \sum_{k=1}^{p} u_k^j \sqrt{\lambda_k} \frac{c^k}{\sqrt{\lambda_k}} = \bar{x}^j + \sum_{k=1}^{p} s_k^j \frac{c^k}{\sqrt{\lambda_k}}$$

donde s_k^j es la saturación del k^{esimo} factor en la variable x^j. Por ejemplo :

$s_1^M = .515 \times \sqrt{28.235} = 2.74$. Así tenemos :

$$M = 9.67 + 2.74 \times F^1 - 1.97 \times F^2$$
$$S = 9.83 + 2.70 \times F^1 - 1.29 \times F^2$$
$$F = 10.20 + 2.62 \times F^1 + 2.28 \times F^2$$
$$L = 10.10 + 2.57 \times F^1 + 1.13 \times F^2$$

donde $F^k = c^k/\sqrt{\lambda_k}$ $k = 1, 2$

II-5. Contribución de las unidades estadísticas a la inercia de la nube

La contribución de la $i^{ésima}$ unidad estadística (u.e.) está dada por :

$$\frac{p_i \, ||x_i'||^2}{\sum_{j=1}^{n} p_j \, ||x_j'||^2}$$

Por ejemplo la contribución de JUAN es :

$$\frac{\frac{1}{9}(6-9.67)^2 + (6-9.83)^2 + (5-10.2)^2 + (5.5-10.1)^2}{40.350} = 0.2104$$

Se calculan estas contribuciones para saber cuales unidades estadísticas son muy diferentes de las otras (no comparables a las otras). Así tenemos que las contribuciones de los individuos son :

.../...

JUAN	.210
ALAN	.0422
ANA	.0679
DIDIER	.125
ANDRES	.0922
MONICA	.269
PEDRO	.115
BRIGITTE	.0593
EVELYNE	.0263

Nota :

Cuando algunas u.e. tienen una contibución muy grande, puede ser útil no hacerlas participar a la dertiminación de las direcciones principales afectándoles un peso nulo. Haremos esta operación en el segundo ejemplo del parágrafo C.

II-6. Contribución de las u.e. a la determinación de los factores principales

La contribución de la i^{ava} u.e. a factor c^k esta dada por :

$$\frac{p_i\ (c_i^k)^2}{\sum_{j=1}^{n} p_j (c_j^k)^2} = \frac{p_i\ (c_i^k)^2}{\lambda_k}$$

Por ejemplo la contribución de JUAN a c^1 es de :

$$\frac{\frac{1}{9}\ (-8.612)^2}{28.235} = .29186 \quad (1)$$

Las contribuciones están dadas en la tabla 4.

.../...

(1) El valor : -8.612 se encuentra en la tabla 5

	1	2
JUA	.292	.0183
ALA	.0592	.0023
ANA	.0406	.111
DID	.162	.0378
AND	.0362	.224
MON	.382	.0033
PIE	.00414	.376
BRI	.015	.163
EVE	.00946	.0641

Tabla 4

II-7. Los factores principales

De la teória sabemos que los factores principales son los vectores proprios de la matriz $W D_p$ donde $W = {}^t X M X$ y que los valores proprios son los mismos de V M (i.e. $Var(c^k) = \lambda_k$).

Los valores de los factores principales para cada individuo están dados en el cuadro siguiente (Tabla 5).

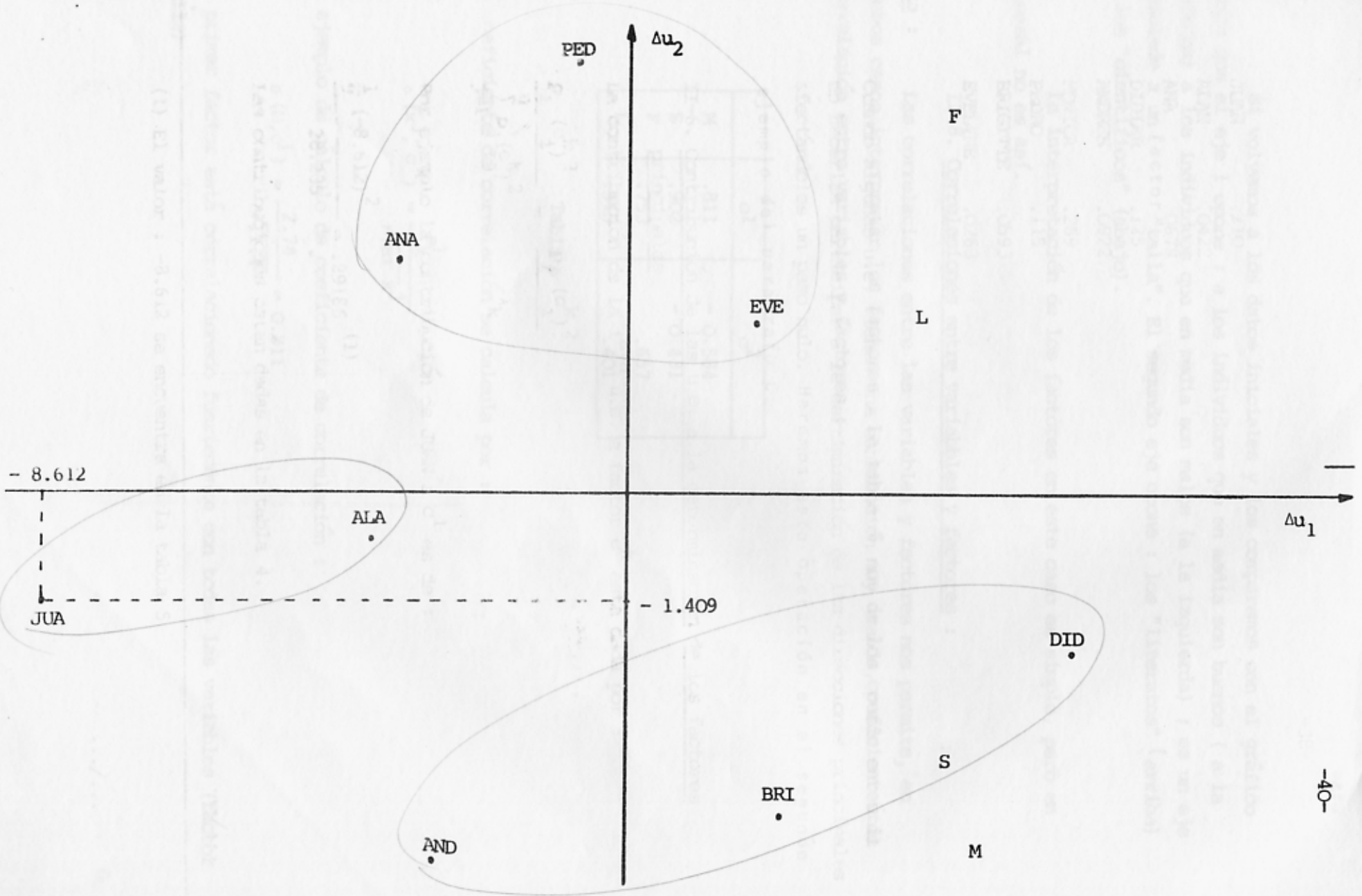
	c^1	c^2
JUA	-8.612	-1.409
ALA	-3.879	-0.522
ANA	-3.213	3.468
DID	6.407	-2.047
AND	-3.033	-4.921
MON	9.852	0.5995
PED	-1.025	6.377
BRI	1.954	-4.200
EVE	1.550	2.634

Tabla 5

En la figura 1 se representa a los individuos tabulados en la tabla 5.

.../...

FIGURA 1


PED
Δu_2
F
ANA
EVE
L
- 8.612
ALA
Δu_1
JUA
- 1.409
DID
S
BRI
AND
M

Si volvemos a los datos iniciales y los comparamos con el gráfico vemos que el eje 1 opone : a los individuos que en media son buenos (a la derecha) a los individuos que en media son malos (a la izquierda) ; es un eje asociado a un factor "talla". El segundo eje opone : los "literatos" (arriba) a los "científicos" (abajo).

La interpretación de los factores en este caso es simple, pero en general no es así.

II-8. Correlaciones entre variabbles y factores :

Las correlaciones entre las variables y factores nos permite, en muchos casos ,interpretar los factores . La tabla 6 nos da los coeficientes de correlación entre variables y factores :

	c^1	c^2
M	.811	- 0.584
S	.902	- 0.431
F	.753	.657
L	.915	.401

Tabla 6

El coeficiente de correlación se calcula por :

$$\rho(x^j, c^k) = \frac{s^j_k}{\sqrt{\text{Var } x^j}}$$

Un ejemplo de cálculo de coeficiente de correlación :

$$\rho(M, c^1) = \frac{2.74}{\sqrt{11.4}} = 0.811$$

El primer factor está correlacionado fuertemente con todas las variables (factor talla)

.../...

El segundo factor está correlacionado positivamente con las variables "literarias" y negativamente con las variables "científicas". Este factor es un factor de "forma".

II-9. Conclusión

En resumen, el análisis en factores principales nos da una "escala" de clasificación de los individuos :

por un lado cuanto más un individuo tenga valores grandes en la dirección del primer factor principal, sus clasificaciones en general son mejores.

por otro lado si un individuo se encuentra en la parte superior (resp. parte inferior) del semiplano de la figura 1 sus calificaciones literarias (resp. científicas) son mejores que las otras.

———— + ————

Talla: fuerte correlación con todos las variables

Forma: correlación positiva con algunas variables
" negativa "

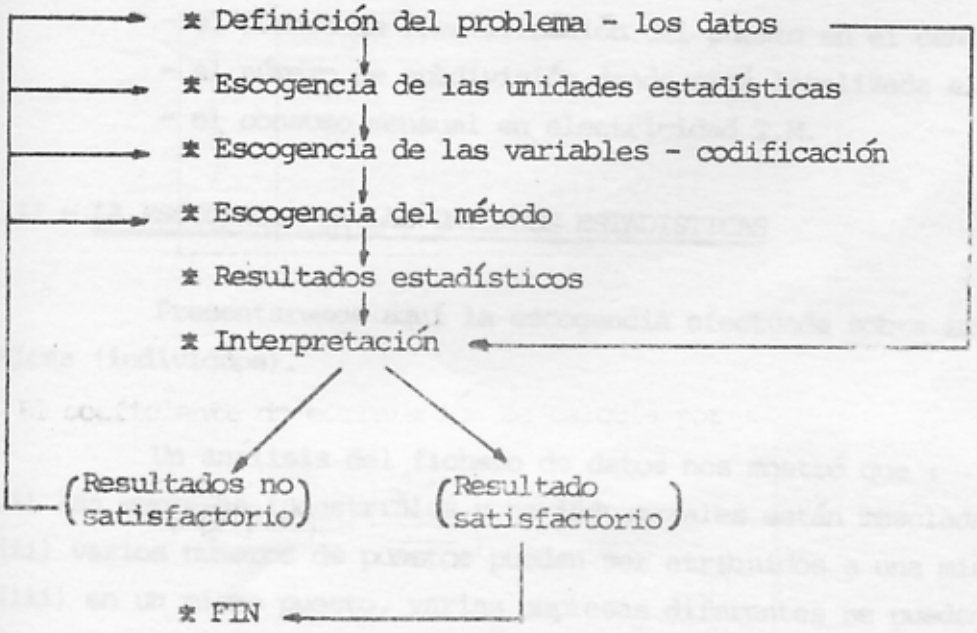
.../...

SEGUNDO EJEMPLO DE ANALISIS EN FACTORES PRINCIPALES :
ESTUDIO DESCRIPTIVO DEL DINAMISMO DE EMPRESAS INDUSTRIALES (3)

I - INTRODUCCION

El estudio de la dinámica de una población de empresas industriales se ha hecho a partir de un índice sobre el consumo mensual de electricidad, utilizando el análisis en componentes principales. Los datos fueron proporcionados por "l'Electricité de France" (E.D.F.).

En el estudio se mostrarán todas las etapas características de una investigación de esta naturaleza. Se procede esquemáticamente de la forma siguiente :



.../...

II - DEFINICION DEL PROBLEMA - LOS DATOS

La población estudiada está constituida por las empresas industriales clientes en electricidad de tensión media (T.M.) de un centro de distribución de la E.D.F.

Los objetivos principales fijados por la E.D.F. son :

a) Definir y adaptar una política tomando en cuenta a los clientes (adecuación de contratos, acción promocional, etc...) ;

b) Proporcionar informaciones conyunturales de los industriales (dinamismo relativo de la empresa en su rama de actividad, dinamismo de las ramas de actividad,...) ;

El centro de la E.D.F. dispone en cada puesto las informaciones siguientes :

- el número de identificación del puesto en el centro,
- el número de subdivisión donde está localizada el puesto,
- el consumo mensual en electricidad T.M.

III - LA ESCOGENCIA DE LAS UNIDADES ESTADISTICAS

Presentaremos aquí la escogencia efectuada sobre las unidades estadísticas (individuos).

Un análisis del fichero de datos nos mostró que :
(i) las empresas industriales y no industriales están mezcladas,
(ii) varios números de puestos pueden ser atribuidos a una misma empresa,
(iii) en un mismo puesto, varias empresas diferentes se pueden suceder en el tiempo,
(iv) ciertas empresas producen una parte de la electricidad que consumen .

En consecuencia, se eliminaron tres tipos de empresas :
a) eliminación de las empresas no industriales,
b) eliminación de los puestos en los cuales se sucedieron varias empresas,
c) eliminación de las empresas auto-productoras de electricidad.

.../...

IV - LOS DIFERENTES ESTUDIOS

Las variables proporcionados por la E.D.F. son el consumo mensual de electricidad. En lo sucesivo se nota i (i=1,n) un puesto de consumo en el estudio, $x_i(\tau)$ el consumo de electricidad T.M. del puesto i en el mes τ ($\tau \in \{1,2,\ldots, p\}$).

En el estudio se tiene p = 36 con τ = 1 para julio de 1974 y τ = p para junio de 1977 y n = 472.

IV-1. Estudio 1 : Descripción por el análisis en componentes principales

NOs proponemos eliminar la parte de las variaciones de los $\{x(\tau) / \tau = 1,\ldots, p\}$ debido a factores "parásitos" en el estudio. Con una recodificación de las variables intentaremos :

i) eliminar el efecto estación, pues depende de la rama de actividad,

ii) poder comparar los dinamismos de puestos teniendo niveles de consumo muy diferentes.

Se trata de satisfacer :

- el objetivo (i) pasando del consumo, a las medias móbiles sobre los 12 meses :

$$m(\tau) = \frac{1}{12} \sum_{k=\tau-6}^{\tau+5} x(k) \qquad \tau \in \{7,\ldots, p-5\}$$

- el objetivo (ii), tomando el crecimiento relativo de estas medias móbiles (expresadas en porcentages) :

$$a(\tau) = 100 \times \frac{m(\tau) - m(\tau-1)}{m(\tau-1)} \qquad \tau \in \{8,\ldots, p-5\}$$

El modelo de A.C.P. nos lleva a las especificaciones siguientes :

- las variables $a(\tau)$, $\tau \in \{8,\ldots, p-5\}$ están centradas
- al puesto i, le corresponde el vector $a_i = {}^t(a_i(8),\ldots, a_i(p-5))$
- el espacio de individuo (R^{24}) está provisto de la métrica euclideana usual
- el espacio de variables (R^{472}) está provisto de la métrica de pesos : se da a cada puesto un peso igual a $\frac{1}{472}$

.../...

Los resultados :

Los porcentajes de inercia explicados por los tres primeros factores principales estan dados en la tabla 1

	Factor 1	Factor 2	Factor 3
% de inercia explicados	24.1	12.7	10.6
% acumulados	24.1	36.8	47.4

Tabla 1

Los valores relativamente pequeños de inercia nos conducen a reconsiderar la escongencia de la codificación y buscar fuentes de dispersión parásitas.

Se pueden hacer dos constataciones :

a) la fórmula de porcentajes anula en parte el efecto de las medias móbiles ; en efecto si simplificamos se tiene :

$$a(\tau) = 100 \times \frac{x(\tau+5) - x(\tau-7)}{m\ (\tau-1)}$$

b) una lectura atenta de los datos dá una anomalía en cuanto a las lecturas no efectuadas. La fórmula precedente muestra que estas anomalías son responsables de variaciones importantes de los valores a (τ) . Este fenómeno es más grave cuanto más estas anomalías no son repartidas al azar ; se observa esto en ciertos meses precisos del año (febrero por ejemplo).

Las figuras 1 , 2-a, 2-b y 2-c confirman el hecho a priori de que la nube de los puestos tiene una forma de "bola" y que sería muy difícil de hacer una clasificación automática : en este caso una descripción con el A.C.P. parece una solución mejor.

.../...

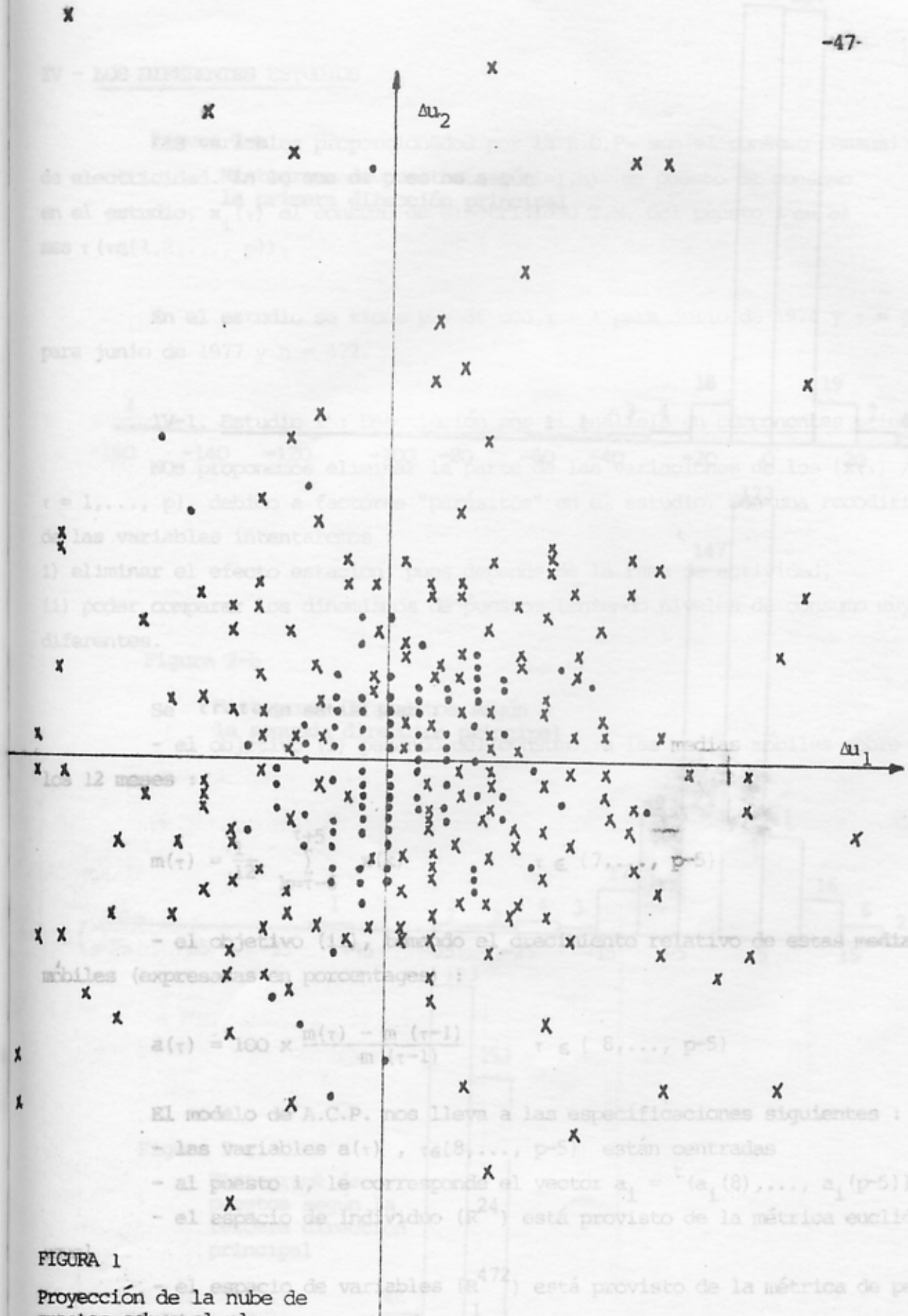


FIGURA 1

Proyección de la nube de puestos sobre el plano principal

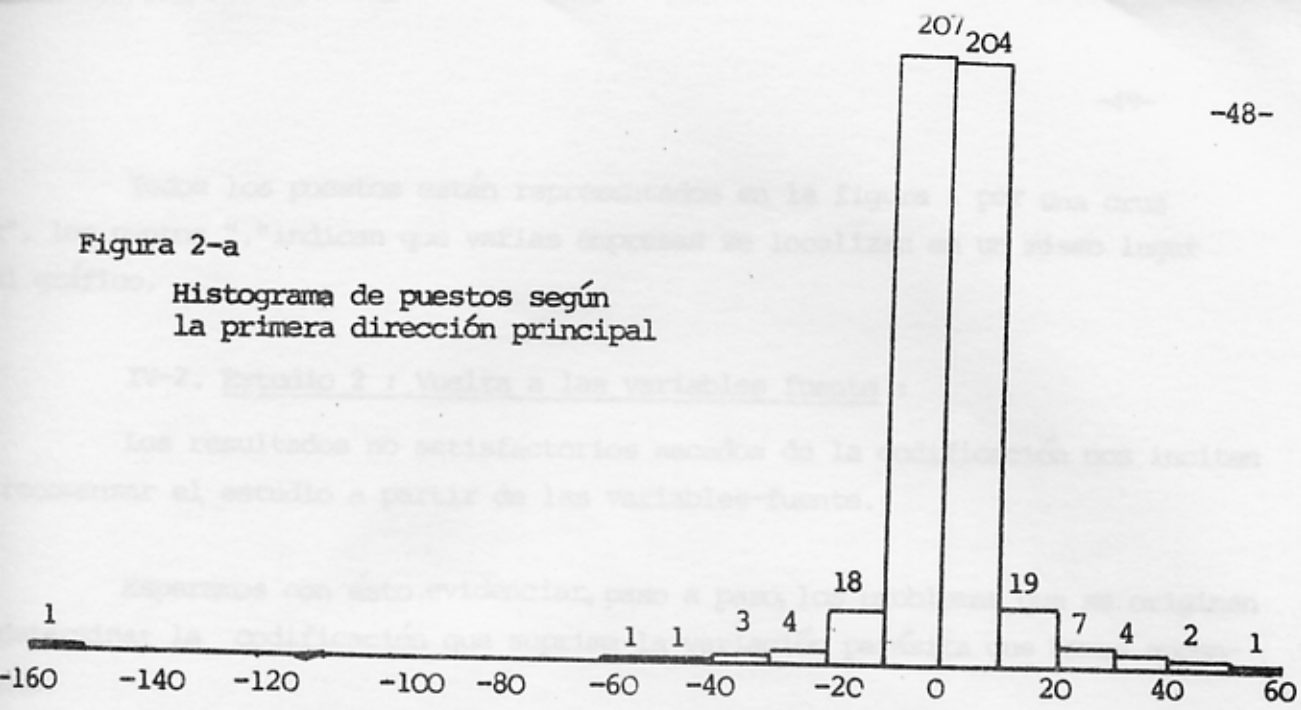


Figura 2-a

Histograma de puestos según la primera dirección principal

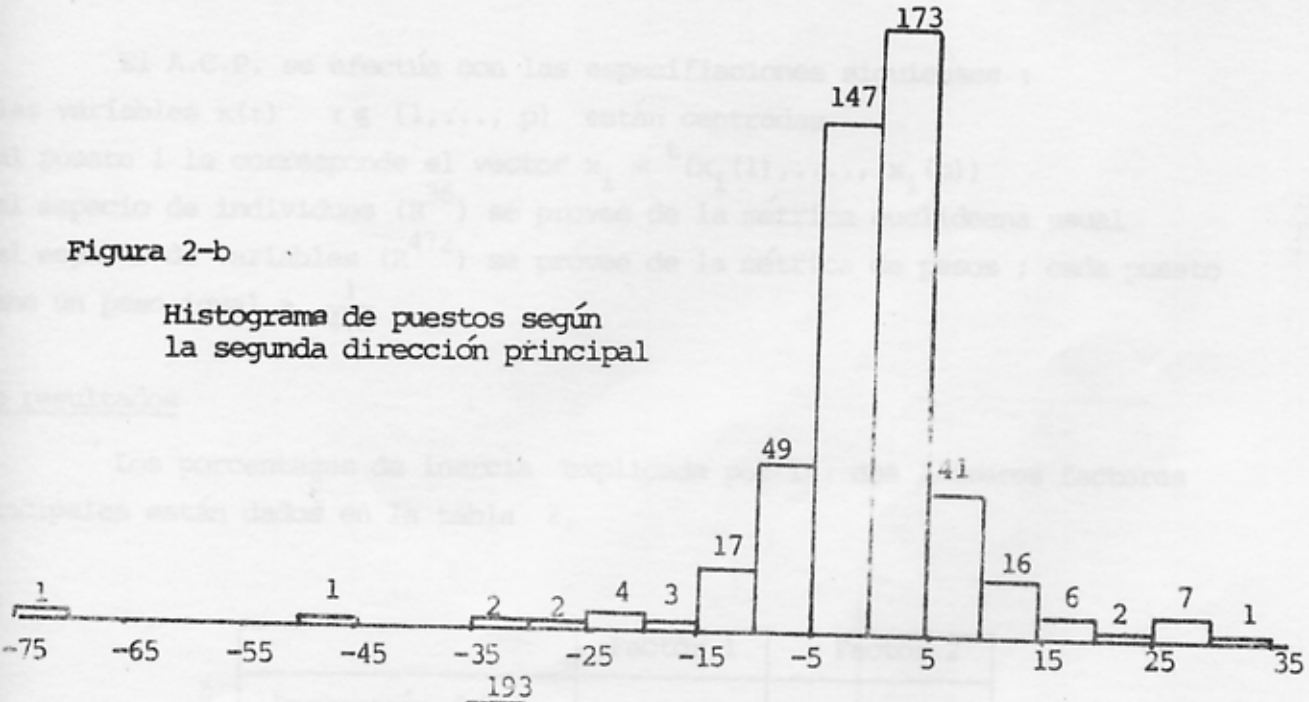


Figura 2-b

Histograma de puestos según la segunda dirección principal

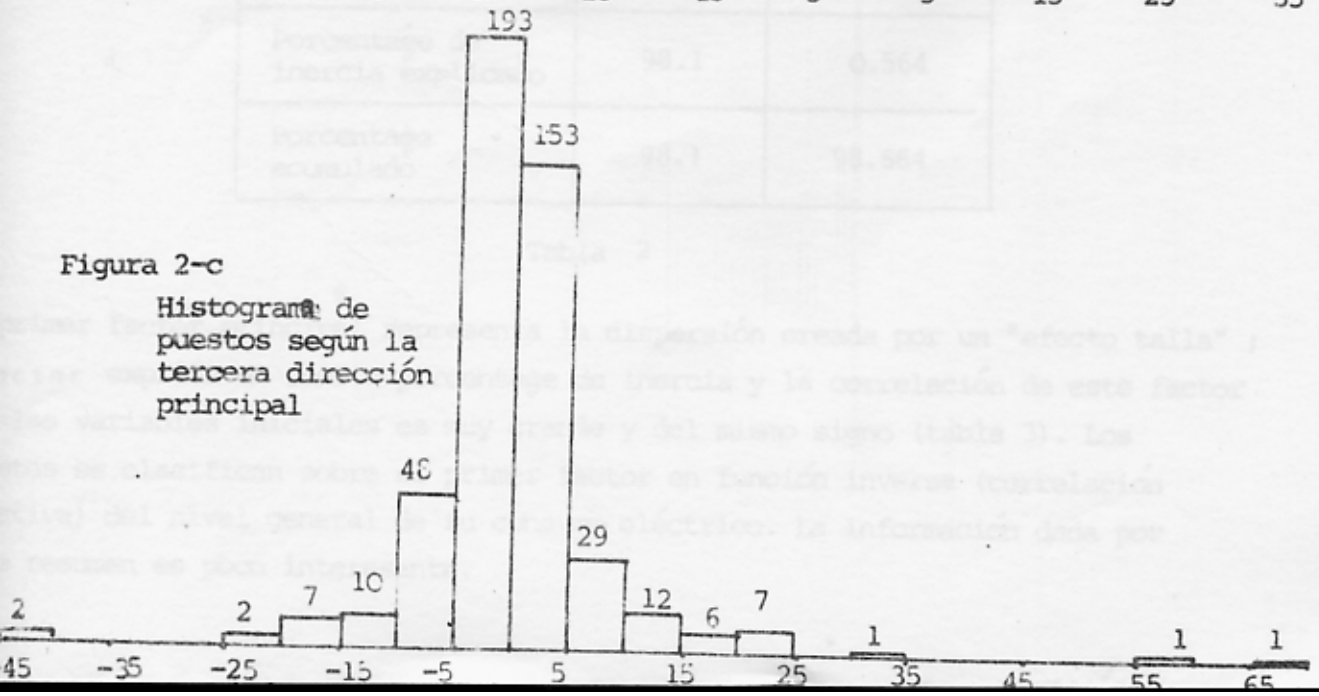


Figura 2-c

Histograma de puestos según la tercera dirección principal

Todos los puestos están representados en la figura 1 por una cruz "x", los puntos "."indican que varias empresas se localizan en un mismo lugar del gráfico.

IV-2. Estudio 2 : Vuelta a las variables fuente :

Los resultados no satisfactorios sacados de la codificación nos incitan a recomenzar el estudio a partir de las variables-fuente.

Esperamos con ésto evidenciar, paso a paso, los problemas que se originan y determinar la codificación que suprima la variación parásita que hemos encontrado.

El A.C.P. se efectúa con las especifiaciones siguientes :
- las variables $x(\tau)$ $\tau \in \{1,\ldots, p\}$ están centradas
- al puesto i le corresponde el vector $x_i = {}^t(x_i(1),\ldots, x_i(p))$
- el espacio de individuos (R^{36}) se provee de la métrica euclideana usual
- el espacio de variables (R^{472}) se provee de la métrica de pesos ; cada puesto tiene un peso igual a $\frac{1}{472}$.

Los resultados

Los porcentages de inercia explicada por los dos primeros factores principales están dados en la tabla 2.

	Factor 1	Factor 2
Porcentage de inercia explicado	98.1	0.564
Porcentage acumulado	98.1	98.664

Tabla 2

El primer factor principal representa la dispersión creada por un "efecto talla" ; el factor explica un fuerte porcentage de inercia y la correlación de este factor con las variables iniciales es muy grande y del mismo signo (tabla 3). Los puestos se clasifican sobre el primer factor en función inversa (correlación negativa) del nivel general de su consumo eléctrico. La información dada por este resumen es poco interesante.

.../...

El segundo factor principal representa la dispersión creada por el rol jugado en los meses de agosto en la actividad industrial : correlaciones (**) más grandes que las otras.

CORRELACIONES DE LAS VARIABLES CON LOS DOS PRIMEROS FACTORES

AÑO	MES	τ	FACTOR 1	FACTOR 2
1974	Julio	1	- 0.991	- 0.063
	Agosto	2	- 0.926 (*)	- 0.313 (**)
	Septiembre	3	- 0.991	- 0.043
	Octubre	4	- 0.992	- 0.082
	Noviembre	5	- 0.986	0.114
	Deciembre	6	- 0.995	- 0.024
1975	Enero	7	- 0.974	- 0.061
	Febrero	8	- 0.992	0.047
	Marzo	9	- 0.995	0.036
	Abril	10	- 0.995	- 0.020
	Mayo	11	- 0.995	- 0.035
	Junio	12	- 0.987	- 0.090
	Julio	13	- 0.980	- 0.100
	Agosto	14	- 0.801 (*)	0.563 (**)
	Septiembre	15	- 0.995	0.009
	Octubre	16	- 0.984	0.019
	Noviembre	17	- 0.994	- 0.014
	Deciembre	18	- 0.996	0.021
1976	Enero	19	- 0.995	0.023
	Febrero	20	- 0.995	0.007
	Marzo	21	- 0.997	0.040
	Abril	22	- 0.995	0.002
	Mayo	23	- 0.990	0.003
	Junio	24	- 0.996	0.019
	Julio	25	- 0.994	0.019
	Agosto	26	- 0.869 (*)	- 0.449 (**)
	Septiembre	27	- 0.995	0.039
	Octubre	28	- 0.997	0.030
	Noviembre	29	- 0.994	0.005
	Deciembre	30	- 0.995	- 0.036
1977	Enero	31	- 0.994	0.024
	Febrero	32	- 0.995	0.036
	Marzo	33	- 0.996	0.055
	Abril	34	- 0.991	0.029
	Mayo	35	- 0.989	- 0.021
	Junio	36	- 0.992	0.075

TABLA 3

Es interesante de notar que el primer factor representa también un poco (correlaciones (*) mas pequeñas que las otras) de la dispersión creada por el rol jugada en las meses de agosto.

.../...

IV-3. Estudio 3 : Atenuación del efecto de estación

El objetivo perseguido es doble : se quiere suprimir las fuentes de variación debidas a los efectos de estación y disminuir las variaciones creadas por la ausencia de lecturas.

Así volvemos a las medias móviles no centradas $m(\tau), \tau=5,\ldots,p-5$. calculadas sobre las variables fuente, las cuales atenúan o suprimen las fluctuaciones estacionarias. Es natural calcular estas medias sobre un período de 12 meses, pues corresponde al intervalo de tiempo de los fenómenos económicos.

Los resultados

El A.C.P. se efectuó con las especificaciones siguientes :

- las variables $m(\tau), \tau\in\{7,\ldots, p-5\}$ están centradas .
- el espacio de las unidades estadísticas (R^{25}) está provisto de la métrica euclideana usual .
- el espacio de variables (R^{472}) está provisto de la métrica de pesos (cada peso es igual a $\frac{1}{472}$).

El porcentage de inercia explicado por el primer factor principal es igual a 99.8 %, y el del segundo factor, que representa en el estudio 2 la fluctuaciones estacionarias, es con esta codificación muy pequeña. Como anteriormente este factor representa la dispersión creada por el efecto "talla". El gráfico de las coordenadas del primer factor (figura 3) muestra que se ha eliminado el efecto de estación. (Es importante de notar que las meses a la izquierda y a la derecha son utilizados por calcular las medias móviles en el estudio 3). A fin de evidenciar las dispersiones (debidas al dinamismo) que deseamos analizar es lógico proponer (en este momento del estudio) una codificación que elimine lo mejor possible el efecto "talla".

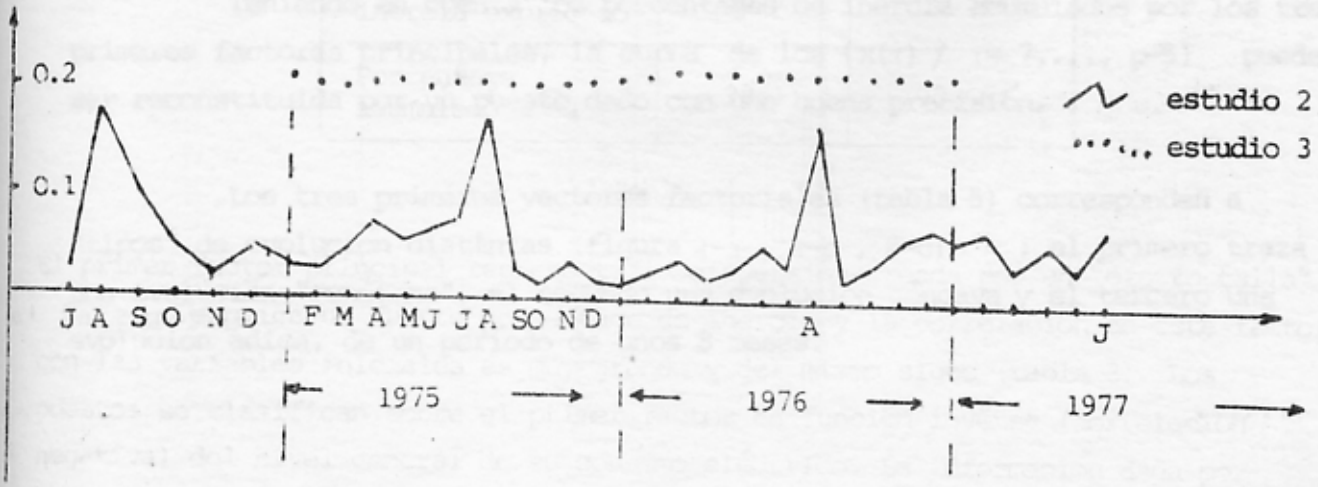


FIGURA 3 : Representación gráfica de las coordenades del primer vector axial factorial

IV-4 Estudio 4 : Eliminación del efecto "talla"

Con el fin de eliminar el efecto "talla" escogimos la codificación $\{ c(\tau) = m(\tau)/ \bar{x} ; \tau=7,..., p-5\}$ donde $\bar{x}$ es la media de consumo sobre los p meses estudiados.

Los resultados

El A.C.P. se efectúa con las especificaciones siguientes :

- las variables $c(\tau)$, $\tau = 7,..., p-5$ están centradas
- el espacio de unidades estadísticas ($\mathbb{R}^{25}$) está provisto de la métrica euclideana usual
- el espacio de variables ($\mathbb{R}^{472}$) se provee de la métrica de pesos (cada peso es igual a $\frac{1}{472}$).

La tabla 4 presenta los porcentajes de inercia explicados por los tres primeros factores principales.

	Factor 1	Factor 2	Factor 3
Porcentage de inercia explicado	66.2	23.5	
Porcentage acumulado	66.2	89.7	3.85

Tabla 4

Teniendo en cuenta los porcentajes de inercia acumulados por los tres primeros factores principales, la curva de los $\{x(\tau) / \tau = 7,..., p-5\}$ puede ser reconstituida por un puesto, dado con una buena precisión.

Los tres primeros vectores factoriales (tabla 5) corresponden a "tipos" de evolución distintas (figura 4-a , 4-b , 4-c) ; el primero traza una evolución "monótona", el segundo una evolución cóncava y el tercero una evolución ádica, de un período de unos 8 meses.

.../...

COORDENADAS DE LOS VECTORES AXIALES FACTORIALES

τ	1er vector	2do vector	3er vector
7	.254	.191	.320
8	.250	.157	.303
9	.247	.136	.263
10	.248	.092	.183
11	.247	.042	.086
12	.238	-.036	-.037
13	.222	-.094	-.107
14	.199	-.147	-.156
15	.163	-.207	-.193
16	.137	-.243	-.183
17	.105	-.282	-.172
18	.082	-.305	-.145
19	.049	-.324	-.047
20	.017	-.325	.032
21	-.011	-.328	.138
22	-.047	-.311	.201
23	-.083	-.278	.271
24	-.135	-.214	.315
25	-.174	-.156	.312
26	-.211	-.101	.264
27	-.234	-.045	.187
28	-.256	.003	.100
29	-.276	.050	.000
30	-.295	.092	-.1 2
31	-.304	.133	-.

TABLA 5

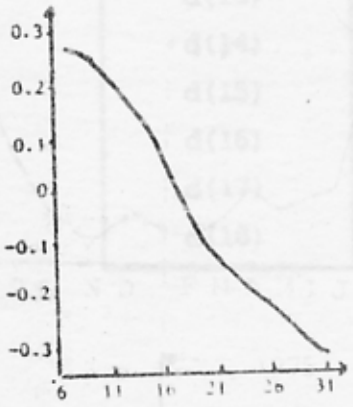


Figura 4-a
1er vector

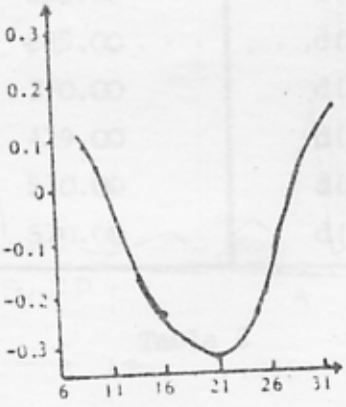


Figura 4-b
2do vector

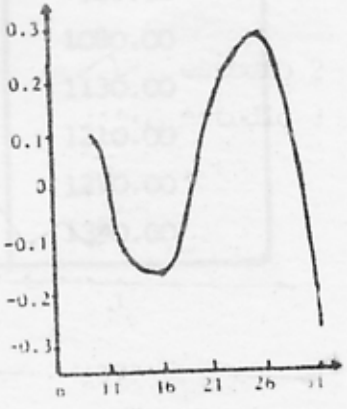


Figura 4-c
3er vector

.../...

IV-5. Estudio 5 : Eliminación de la dispersión creada por la ausencia de un origen común.

Podemos aún mejorar la codificación volviendo a un origen común las "tazas de dinamismo". Suprimimos así una fuente de dispersión no interesante al estudio.

La codificación propuesta es :

$$d(\tau) = 100 \times \frac{m(\tau) - m(7)}{\bar{x}} \quad , \tau = 7,\ldots, p-5$$

Es natural observar (tabla 6) que las varianzas de las variables $d(\tau)$ crecen con τ.

Variables	Varianza	Variables	Varianza
d(7)	0	d(19)	629.00
d(8)	17.70	d(20)	677.00
d(9)	34.90	d(21)	733.00
d(10)	70.20	d(22)	789.00
d(11)	116.00	d(23)	840.00
d(12)	179.00	d(24)	908.00
d(13)	230.00	d(25)	983.00
d(14)	285.00	d(26)	1080.00
d(15)	370.00	d(27)	1130.00
d(16)	429.00	d(28)	1210.00
d(17)	510.00	d(29)	1290.00
d(18)	570.00	d(30)	1390.00

Tabla 6

.../...

Sabemos que si el espacio de individuos se provee de la métrica euclideana usual, las variables tienen un peso en el A.C.P.mayor, cuanto mayor sea su varianza. La escogencia del origen en τ =7 permite de dar una mayor importancia a las variables que miden el dinamismo más recientemente. Si se escoge el origen en τ= 19 se observan varianzas decrecientes hasta en τ=19 y después varianzas crecientes hasta en τ=31. Para esta escogencia se había dado más importancia al pasado lejano que al reciente.

IV-5.1. Resultados del primer A.C.P.

El A.C.P. se efectua con las espicificaciones siguientes :

- las variables d(τ) están centradas
- el espacio de individuos (R^{24}) se provee de la métrica euclideana usual
- el espacio de variables (R^{472}) se provee de la métrica de pesos (cada peso es igual a $\frac{1}{472}$).

El porcentage de inercia explicado por los dos primeros factores prin - cipales es mejor que en el estudio precedente (tabla 7)

	Factor 1	Factor 2
% de inercia explicado	83.1	12.6
% acumulado	83.1	95.7

Tabla 7

Eso es natural porque la codificacion ha eliminado (en la nube) las variaciones creadas por la ausencia de un origen común. Relativamente al objetivo del estudio el resumen estadística es mejor.

.../...

Si analizamos de cerca los resultados se observa que 15 puestos son responsables del 34 % de la inercia total. En este contexto, corremos el riesgo de que los resultados obtenidos dependan "mucho" de los tipos de dinamismo de estas 15 empresas, pues representan el 3 % de las empresas estudiadas. Una solución razonable consiste en afectar un peso nulo a estas empresas y hacer un nuevo análisis.

IV-5.2. Resultados del segundo A.C.P.

La única diferencia en las especificaciones reside en la métrica del espacio de variables (R^{472}):

(i) un peso nulo es dado a los 15 puestos que tienene una fuerte contribución a la inercia total
(ii) un peso igual a $\frac{1}{457}$ es dado a los otros.

Con respecto al A.C.P. anterior se observa una estabilidad de los resultados pero con una redistribución de la inercia (tabla 8)

	Factor 1	Factor 2
% de inercia explicada	78.6	16.0
% acumulado	78.6	94.6

Tabla 8

Como en el análisis precedante los 3 primeros vectores principales corresponden a los tres tipos de evolución (figura 4a,b,c). Es importante hacer notar que es posible dar rápidamente (parágrafo VIII de las bases matemáticas del A.C.P.) un diagnóstico sobre la dinámica de una empresa por la posición en el plano principal (figura 5).

.../...

Hemos hecho figurar las curvas de "tazas de dinamismo" de ciertas empresas en círculos y las curvas tipos correspondientes a u_1, $u_1 + u_2$, u_2, $u_2 - u_1$, $-u_1 - u_2$, $-u_2$, $-u_2 + u_1$ en rectángulos.

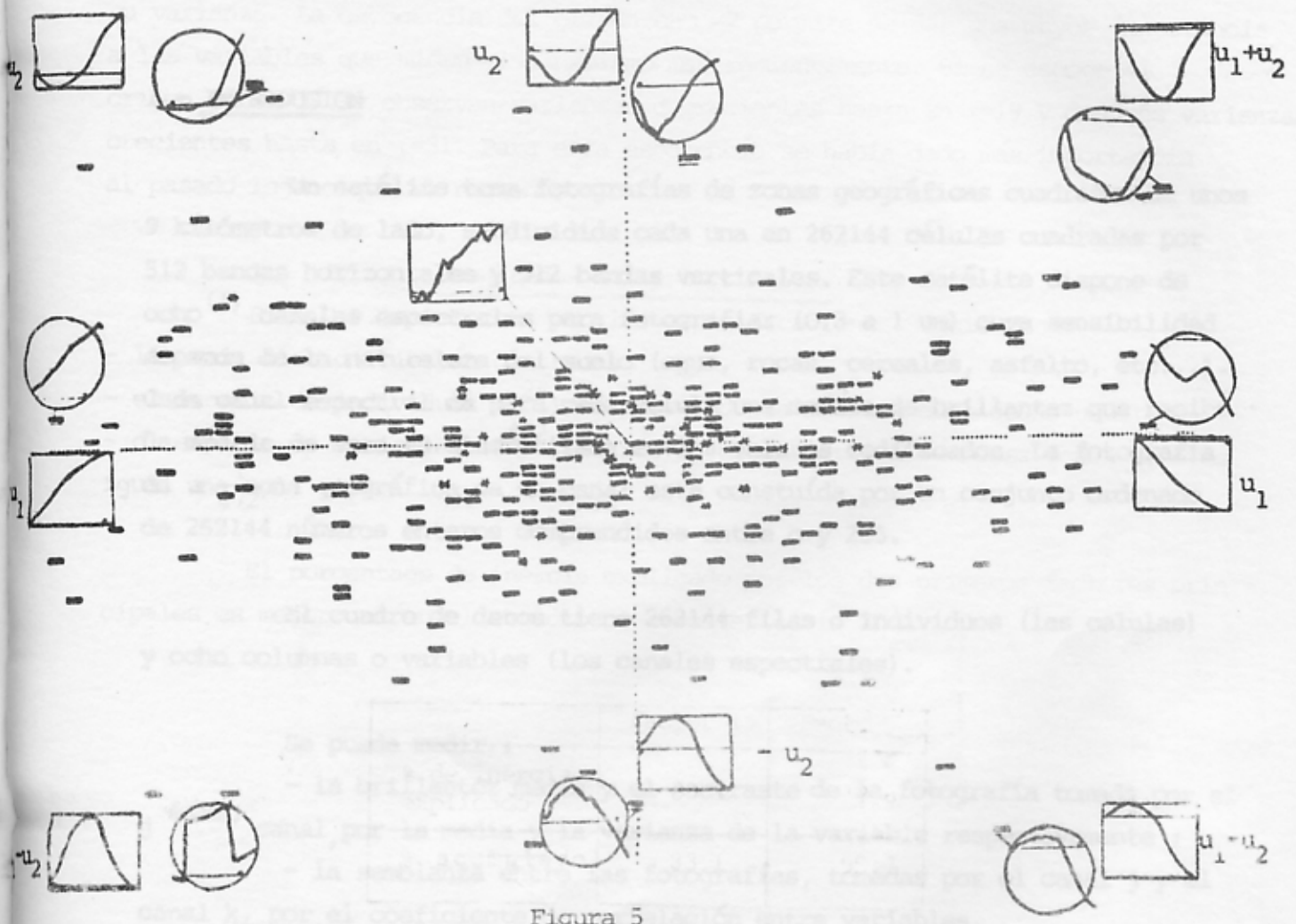


Figura 5

IV-6. Conclusión

En el A.C.P. lo que se busca es ANALIZAR LA DISPERSION que nos interesa. Así cuando en un problema dado aparecen dispersiones que perturban muy poco el estudio en el A.C.P, igualmente nos da una información valiosa del objeto analizado. Pero si por el contrario las dispersiones que perturban el fenómeno son más grandes que aquellas que se quieren analizar (como es el estudiado) se hace necesario buscar una CODIFICACION que elimine estas dispersiones "parásitas" del objeto de estudio.

Hemos hecho figurar aquí todas las etapas con fines pedagógicos.

TERCER EJEMPLO : ANALISIS EN FACTORES PRINCIPALES Y EN FACTORES SEMEJANTES

PROBLEMAS QUE SE PRESENTAN EN LA COMPRESIÓN Y LA TELETRANSMISIÓN DE IMAGENES MULTIESPECTRALES (7)

I - INTRODUCION

Un satélite toma fotografías de zonas geográficas cuadradas de unos 9 kilómetros de lado, subdividida cada una en 262144 células cuadradas por 512 bandas horizontales y 512 bandas verticales. Este satélite dispone de ocho[1] canales espectrales para fotografiar (0.3 a 1 um) cuya sensibilidad depende de la naturaleza del suelo (agua, rocas, cereales, asfalto, etc...). Cada canal espectral da para cada celula una medida de brillantez que recibe · La medida de esta es discretizada en 256 valores codificados. La fotografía de una zona geográfica de un canal esta constuída por un conjunto ordenado de 262144 números enteros comprendidos entre o y 255.

El cuadro de datos tiene 262144 filas o individuos (las celulas) y ocho columnas o variables (los canales espectrales).

Se puede medir :

- la brillantez media y el contraste de la fotografía tomada por el $j^{ésimo}$ canal, por la media y la varianza de la variable respectivamente ;
- la semblanza entre las fotografías, tomadas por el canal j y el canal k, por el coeficiente de correlación entre variables.

En lo que sigue, damos[2] (tabla 1 y 2) la matriz de varianzas-covarianzas y de correlaciones ; todas las celulas tienen el mismo peso $p_i = \frac{1}{262144}$

.../...

(1) El conjunto constituido por las ocho fotografías se llama imagen multispectral

(2) Los calculos han sido efectuados en el "Centre National d'Etudes Spatiales" Toulouse - FRANCIA

	canal 1	canal 2	canal 3	canal 4	canal 5	canal 6	canal 7	canal 8
canal 1	.280							
canal 2	.326	.409						
canal 3	.397	.491	.628					
canal 4	.452	.552	.682	.798				
canal 5	.450	.554	.694	.793	.829			
canal 6	.301	.384	.494	.528	.532	.565		
canal 7	.150	.207	.277	.259	.244	.533	.747	
canal 8	.110	.152	.204	.194	.185	.408	.569	.447

Tabla 1 : Matriz de varianza-covarianza

	canal 1	canal 2	canal 3	canal 4	canal 5	canal 6	canal 7	canal 8
canal 1	1							
canal 2	.962	1						
canal 3	.948	.968	1					
canal 4	.956	.965	.964	1				
canal 5	.934	.952	.962	.975	1			
canal 6	.758	.798	.829	.787	.777	1		
canal 7	.329	.374	.405	.336	.311	.821	1	
canal 8	.312	.356	.386	.325	.304	.813	.985	1

Tabla 2 : Matriz de correlaciones

Observaciones :

- la fotografía más contrastada fue tomada por el canal 5 (varianza = 0.829) y la menos contrastada por el canal 1 (var .280).

- como es previsible los canales están muy relacionados (las fotografias se asemejan) y se correlación es mayor cuanto más cerca están en el espectro.

II - COMPRESION DE LA IMAGEN MULTIESPECTRAL - APPLICACION DEL A.C.P.

Venimos de constatar (tabla 2) que las informaciones dadas por los ocho canales son redundantes (las correlaciones grandes). Con el fin de no saturar los canales de transmision - satélite a tierra y viceversa - con informaciones inútiles, se comprimen (resumen) las informaciones en un computador a bordo del satélite. Esta operación se realiza por medio del análisis en componentes principales (A.C.P.). Para hacer que cada canal juegue un rol más importante cuando la fotografía es más contrastada (varianza del canal espectral grande) las variables no se reducen. La distancia utilizada es la usual (la base canónica de E es M ortogonal) y todas las células tienen el mismo peso $\frac{1}{262144}$. El A.C.P. aplicado a la imagen multiespectral da los resultados en las tablas 3 y 4.

.../...

	FACTORES PRINCIPALES							
	1	2	3	4	5	6	7	8
% inercia explicado	75.51	22.33	0.79	0.51	0.29	0.24	0.18	0.15
% acumulado	75.51	97.84	98.63	99.14	99.43	99.67	99.85	100.00

Tabla 3

En la tabla 3 los porcentages representan la parte de la varianza total (1) reproducida explicada por el factor de la columna correspondiente.

	FACTORES PRINCIPALES						
	c^1	c^2	c^3	c^4	c^1,c^2	c^1,c^2,c^3	c^1,c^2,c^3,c^4
canal 1	.856	0.078	0.036	0.006	0,934	0,970	0,976
canal 2	0.905	0.058	0.019	3.10^{-3}	0,962	0,981	0,981
canal 3	0.930	0.043	0.003	0.019	0,973	[illegible]	0,995
canal 4	0.902	0.080	[illegible]	0.012	0,982	0,982	0,994
canal 5	0.864	0.093	0.038	9.10^{-3}	0,977	0,995	0,995
canal 6	0.870	0.116	0.003	0.001	0,986	0,988	0,989
canal 7	0.351	0.644	0.001	3.10^{-3}	0,994	0,995	0,995
canal 8	0.336	0.651	0.001	0.001	0,988	0,988	0,989

Tabla 4

La tabla 4 da los cuadrados de los coeficientes de correlación, simples y multiples : "canales x factores". Estas cantidades miden también por cada canal espectral la proporción de varianza reproducida por los factores principales.

.../...

(1) Suma de las varianzas de los 8 canales espectrales

III - TELETRANSMISION DE UNA IMAGEN MULTIESPECTRAL COMPRIMIDA - UTILIDAD DEL ANALISIS EN FACTORES SEMEJANTES

III-1. Cualidades deseadas para una transmision optimal : satélite-tierra

Despues del A.C.P. las informaciones deben ser transmitidas a la tierra. Cabe preguntarse : ¿ las informaciones están adaptadas a los riesgos de pérdida de información y a las necesidades (almacenamiento) de la teletransmisión ?

Enunciemos de manera no formal, las cualidades ideales que deban poseer éstas informaciones, afin de realizar una teletransmisión optimal.

a) todas las informaciones características de una misma célula deben ser de igual importancia. Por ejemplo, dado que las informaciones son las coordenadas de un vector y que a priori cada información tiene la misma probabilidad de perderse en la transmisión, parece natural tratar de que cada coordenada aporte la misma información (contribución a la norma del vector).

b) cada información perdida debe ser reconstruida al máximo por las otras informaciones de la célula.

c) para una célula, la precisión con la cual se reconstruye una información perdida es la misma para todas las informaciones.

d) el orden de tamaño de las informaciones, debe ser la misma, con el fin de optimizar las zonas de almacenamiento de información.

Es claro que los factores principales no convienen, pues tienen las cualidades opuestas. Es necesario encontrar cuatro variables nuevas que llamaremos "factores semejantes", combinaciones lineales de los cuatro factores principales, tales que las cualidades deseadas sean respetadas al máximo. Para la resolución del problema pueden referirse a SCHEKTMAN (8).

III-2. Factores semejantes de la imagen multiespectral

Cuando el número de factores principales retenidos es igual a 1, 2 o un múltiplo[1] de 4, existen expresiones simples que dan los factores semejantes en función de los factores principales. En el ejemplo presente los ejes factoriales principales (u_1, u_2, u_3, u_4) están dados. Si escogemos las expresiones, para los cuatro ejes factoriales Semblables (s_1, s_2, s_3, s_4) se obtiene :

.../...

(1) al menos para los valores no mayores de 116.

$$s_1 = \frac{u_1 + u_2 + u_3 + u_4}{2} \quad ; \quad s_2 = \frac{u_1 + u_2 - u_3 - u_4}{2}$$

$$s_3 = \frac{u_1 - u_2 + u_3 - u_4}{2} \quad ; \quad s_4 = \frac{u_1 - u_2 - u_3 + u_4}{2}$$

A las direciones semejantes s le corresponden los factores semejantes S que están conformes a las exigencias enunciadas en el parágrafo III-1. Así tenemos los resultados siguientes :

	S^1	S^2	S^3	S^4
S^1	1.17			
S^2	1.14	1.17		
S^3	0.63	0.62	1.17	
S^4	0.62	0.63	1.14	1.17

Matriz de varianza-covarianza

	S^1	S^2	S^3	S^4
S^1	1.00			
S^2	0.97	1.00		
S^3	0.54	0.53	1.00	
S^4	0.53	0.54	0.97	1.00

Matriz de correlación

$$\text{Var}\,[S^1/S^2, S^3, S^4] = \text{Var}\,[S^2/S^1, S^3, S^4] = \text{Var}\,[S^3/S^1, S^2, S^4] =$$

$$\text{Var}\,[S^4/S^1, S^2, S^3] = 0.057$$

$$\rho[S^1; S^2, S^3, S^4] = \rho[S^2; S^1, S^3, S^4] = \rho[S^3; S^1, S^2, S^4] = \rho[S^4; S^1, S^2, S^3]$$

$$= \sqrt{1 - 0.057 \times \frac{1}{1,17}} = 0.975$$

Es importante de notar que los coeficientes de correlación múltiple son grandes (0.975) y las varianzas condicionadas son relativamente pequeñas (.057);esto permite reconstruir una información perdida con un error pequeño.

En definitiva hemos visto como se han resulto los problemas que se presentaron en el caso estudiado :

1 - comprimir la información (A.C.P.)

2 - Transmitir la información con un riesgo de perdida mínimo (análisis en factores semejantes).

.../...

CUARTO EJEMPLO : ANALISIS DE CORRESPONDENCIAS

LAS CONDICIONES DE VIVIENDA DE LOS TRABAJADORES IMIGRANTES EN TOULOUSE

En esta exposición consideraremos una tabla de contingencia (tabla 1) extraída de un estudio realizado (6) en la Universidad du Mirail (Toulouse) sobre las condiciones de viviendra de los trabajadores emigrandos en Toulouse.

En el estudio consideraremos dos variables cualitativas. La primera variable cualitativa (horizontal) representa el barrio, codificada en 15 modalidades (una por cada barrio que se toma en cuenta en el estudio). La segunda variable cualitativa (vertical) es uno de los indicadores de la encuesta sobre las condiciones de vivienda. Las modalidades son : wc en el apartamento, en el piso, en el edificio, sin wc.

La tabla 1 tiene todas las informaciones útiles para una interpretación (efectivo, porcentage con respecto al total general (P.G.), porcentage lineal (P.L), porcentage columna (P.C.), efectivos teóricos (E.F.T.) y contribución de cada célula de la tabla al valor χ^2 (c.c.x.2.). Se puede ver a simple vista que la tabla 1 posee informaciones interesantes.

Si observamos los valores de P.C. y CCX2 notamos que los barrios (1,2), (5,12,13), (3,9), (7,14) tienen valores muy parecidos lo que nos hace pensar en una cierta proximidad dentro de cada grupo. Pero vemos que este trabajo es fastidioso y difícil, por todas las relaciones que hay que sacar. Y lo sería aún más si el número de modalidades de las variables fuera mayor.

.../...

TABLA 1

BAÑOS / WC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	TOTAL
APARTAMENTO																
EFECTIVO	36.00	24.00	42.00	33.00	110	10.00	19.00	26.00	36.00	51.00	01.00	55.00	06.00	14.00	23.00	486
P.G.	.043	.029	.050	.039	.131	.012	.023	.031	.043	.061	.001	.065	.007	.017	.027	.580
P.L.	.074	.049	.086	.068	.225	.020	.039	.053	.074	.105	.002	.113	.012	.029	.047	
P.C.	.254	.229	.575	.440	.957	.714	.760	.578	.600	.836	.056	1.000	1.000	.667	.920	
E.F.T.	82.00	61.00	42.00	43.00	67.00	08.00	14.00	26.00	33.00	33.00	10.00	32.00	03.00	12.00	14.00	
C.C.X.2.	26.05	22.32	0	2.52	28.19	.44	1.40	0	.04	6.92	8.53	16.77	1.83	.27	5.00	
PISO																
EFECTIVO	34.00	21.00	09.00	04.00	01.00	01.00	01.00	01.00	03.00	05.00	05.00	0	0	0	0	85
P.G.	.040	.025	.011	.005	.001	.001	.001	.001	.004	.006	.006	0	0	0	0	.101
P.L.	.400	.247	.106	.047	.012	.012	.012	.012	.035	.059	.059	0	0	0	0	
P.C.	.239	.200	.123	.053	.009	.071	.040	.022	.050	.082	.278	0	0	0	0	
E.F.T.	14.00	11.00	07.00	08.00	12.00	01.00	03.00	05.00	06.00	06.00	02.00	06.00	01.00	02.00	03.00	
C.C.X.2.	26.98	10.20	.36	1.68	9.70	.12	.92	2.76	1.54	.22	5.58	5.55	.61	2.12	2.52	
EDIFICIO																
EFECTIVO	67.00	59.00	22.00	28.00	04.00	03.00	03.00	16.00	21.00	05.00	02.00	0	0	05.00	0	235
P.G.	.080	.070	.026	.033	.005	.004	.004	.019	.025	.006	.002	0	0	.006	0	.279
P.L.	.285	.251	.094	.119	.017	.013	.013	.068	.089	.021	.009	0	0	.021	0	
P.C.	.472	.562	.301	.373	.035	.214	.120	.356	.350	.082	.111	0	0	.238	0	
E.F.T.	40.00	29.00	20.00	21.00	32.00	04.00	07.00	13.00	17.00	17.00	05.00	15.00	02.00	06.00	07.00	
C.C.X.2.	18.90	30.09	.13	2.39	24.59	.01	2.27	.94	1.08	8.49	1.82	15.35	1.67	.13	6.98	
SIN																
EFECTIVO	05.00	01.00	0	10.00	0	0	02.00	02.00	0	0	10.00	0	0	02.00	02.00	34
P.G.	.006	.001	0	.012	0	0	.002	.002	0	0	.012	0	0	.002	.002	.040
P.L.	.147	.029	0	.294	0	0	.059	.059	0	0	.294	0	0	.059	.059	
P.C.	.035	.010	0	.133	0	0	.080	.044	0	0	.556	0	0	.095	.080	
E.F.T.	06.00	04.00	03.00	03.00	05.00	01.00	01.00	02.00	02.00	02.00	01.00	02.00	0	01.00	01.00	
C.C.X.2.	.09	2.48	2.95	16.05	4.64	.57	.97	.02	2.42	2.46	11831	2.22	.24	1.57	.97	
TOTAL																
EFECTIVO	142	105	73	75	115	14	25	45	60	61	18	55	6	21	25	840
P.G.	.169	.125	.087	.089	.137	.017	.030	.053	.071	.072	.021	.065	.007	.025	.030	

Con el objeto de completar el estudio de la asociación entre las variables (BARRIO x WC) con un resumen más rico y global utilizamos, en este caso[1], el análisis de correspondencias. Este modelo es un caso particular del análisis en factores principales :

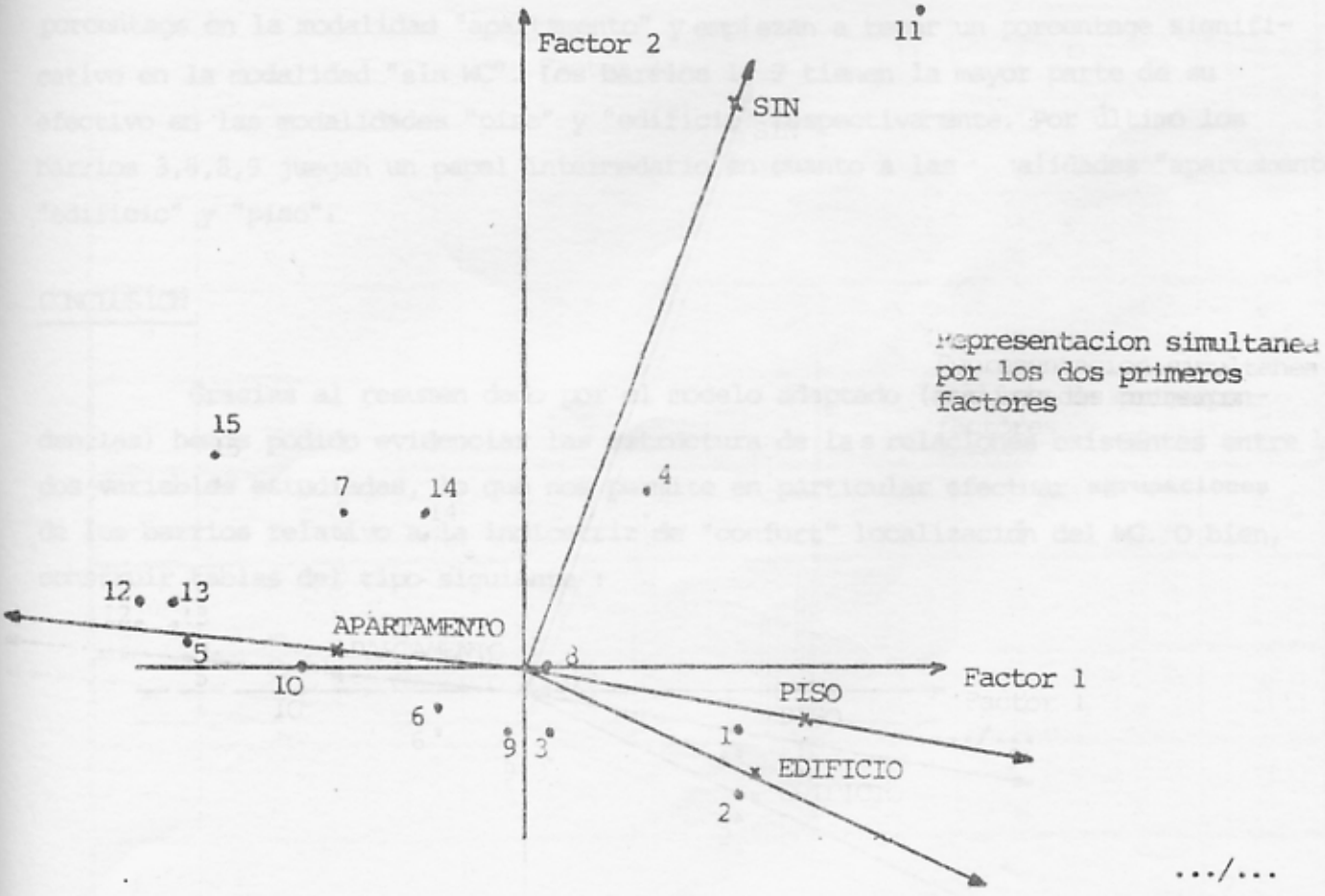
a) los pesos asociados a las unidades estadisticas (barrio[2] por ejemplo) son iguales al total lineal dividido por la talla de la población. Por ejemplo el peso del barrio n°1 es igual a 0.169.

b) la distancia escogida es la del CHI-cuadrado (χ^2).

Observamos en la tabla 2 y el gráfico los principales resultados del análisis de correspondencias de la tabla 1.

	Factor 1	Factor 2	Factor 3
% de inercia explicado	62.55	32.91	4.54
% acumulado	62.55	95.46	100.00

Tabla 2



.../...

(1) Estudio de la relación entre dos variables cualitativas

(2) Se demuestra que se pueden intercambiar los papeles jugados por líneas y columnas en el análisis de correspondencias

Dado que el plano principal explica el 95.46 % de la inercia de la nube la precisión con la que se proyectan los datos es excelente.

Las observaciones hechas anteriormente en la tabla 1 aparecen en la figura , es decir, la proximidad de (7,14), (3,9), (12,13,5) y (1,2). Pero además aparecen claras otras relaciones entre las modalidades de la primera variable (barrios). Por otro lado en el gráfico podemos ver otras relaciones que son importantes : cuanto más un barrio se aleje en la dirección de una modalidad de la segunda variable (WC) tanto mayor es su porcentage del efectivo en esa modalidad.

Así podemos notar que 11 juega un rol particular : es el único barrio que se aleja tanto en la dirección de la modalidad "sin WC" pues el 55 % de su efectivo pertenece a ésta.

De igual manera los barrios 12, 13, 5, 10 poseen la mayor parte de su efectivo en la modalidad "apartamento". Los barrios 7, 14, 15 poseen en alto porcentage en la modalidad "apartamento" y empiezan a tener un porcentage significativo en la modalidad "sin WC". Los barrios 1, 2 tienen la mayor parte de su efectivo en las modalidades "piso" y "edificio" respectivamente. Por último los barrios 3,6,8,9 juegan un papel intermedario en cuanto a las modalidades "apartamento", "edificio" y "piso".

CONCLUSION

Gracias al resumen dado por el modelo adaptado (análisis de correspondencias) hemos podido evidenciar las estructura de las relaciones existentes entre las dos variables estudiadas, lo que nos permite en particular efectuar agrupaciones de los barrios relativo a la indicatriz de "confort" localización del WC. O bien, construir tablas del tipo siguiente :

.../...

PARTE 2

ESTADISTICA

INFERENCIAL

(ANALISIS LINEAL DE DATOS UNI O MULTIDIMENSIONALES)

Barrio / Porcentage de WC en	12	15	13	5	7	10	14
Apartamento	1	.92	1	.957	.76	.836	.667
Edificio	0	0	0	0.035	.120	.082	.238
Piso	0	0	0	0.009	0.04	0.082	0

Barrio / Porcentage de WC en	6	9	8	3	4	1	2
Apartamento	.714	.60	.578	.575	.44	.254	.229
Edificio	.214	.350	.356	.301	.373	.472	.562
Piso	0.071	.C5	0.022	0.123	0.053	.239	.200

Tabla 3

En estas tablas, el orden de los barrios es el que se construye proyectando estos sobre una recta pasando por la direccion "apartamento" y entre las direcciones "edificio" y "piso".

———— + ————

.../...

En la introducción hemos señalado que los cuadros de datos pueden estar estratificados según :

- el conjunto de individuos : análisis de varianza uni o multivariable. Es el caso de los modelos a efectos fijos, con efectos aleatorios y el modelo mixto;

- el conjunto de variables : análisis de regresión sobre variables explicativas con o sin error;

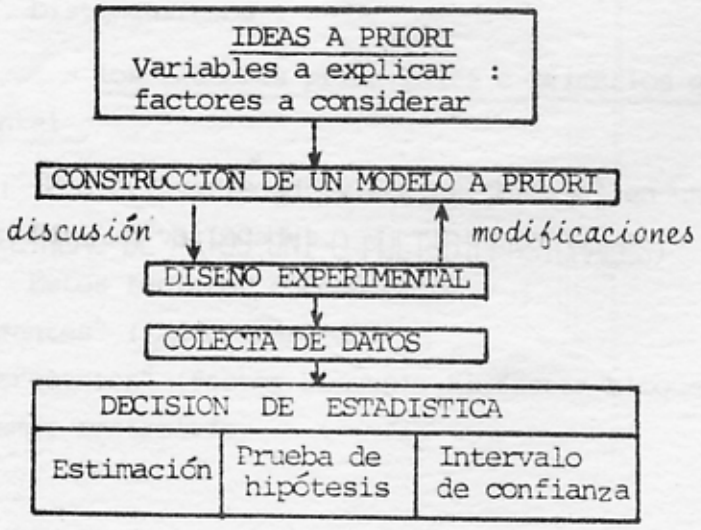
- el conjunto de las variables y de los individuos : análisis de covarianza.

Hemos de hacer notar que todos estos modelos caen dentro del marco de lo que se llama modelo lineal general.

Según nuestro parecer la práctica de estos modelos conllevan tres fases que son las siguientes :

- la descripción del fenómeno a estudiar y la concepción de un diseño experimental,
- la realización del diseño experimental,
- el análisis y la interpretación de resultados.

De este modo, las tres fases se pueden resumir en el siguiente esquema :



.../...

Para el estudio de un fenómeno, es necesario escoger las variables a explicar (o predecir). Estas variables se llaman también dependientes ; es el resultado que observamos del fenómeno estudiado. Necesitamos escoger o definir las variables explicativas (o factores) que tienen acción directa sobre el fenómeno.

I - DIFERENTES TIPOS DE FACTORES

Hacemos diferencia entre :

Factores cuantitativos discretizados : aquí nos referimos a los niveles de un factor.

Ejemplo :

Factor "dosis de nitrogeno" en una experiencia de fertilización. Es necesario escoger los niveles en números y posición : (problema de superficie de respuesta).

Factores cualitativos : Hacemos referencia a las modalidades de un factor.

Ejemplo :

Factor "variedad" en una prueba de rendimiento. Es necesario saber el número de modalidades y el número de repeticiones por modalidad.

Distinguiremos :

- los factores principales o primarios que definen el diseño experimental

Ejemplo : Factor "variedad' y factor "bloque" en un diseño de bloques para comparar las variedades.

Estos factores pueden ser :

"interesantes" (factor variedad)

"no interesantes" (factor bloque). El factor bloque es un factor que llamaremos controlado.

.../...

- los factores secundarios : son medidas sobre el diseño experimental. Estas variables "no interesantes" , "no controladas" aparecen como variables concomitantes (covariables) en el modelo de análisis.

- los factores de error : "no controlados", "no medidos", son aleatorios.

II - DIFERENTES TIPOS DE MODELO

Distinguiremos tres tipos de modelos.

Modelo a efectos fijos :

Los niveles o modalidades de cada factor son escogidos y bien definidos.

Modelo a efectos aleatorios :

Las modalidades de los factores son extraídas de las poblaciones correspondientes a los diferentes factores.

Modelo mixto :

Algunos factores son a efectos fijos y los otros a efectos aleatorios.

Una vez que las variables a explicar están escogidas y los factores definidos se contruye un modelo a priori y se escoge un diseño experimental apropriado. Es el trabajo que debe hacer en común el experimentador y el estadístico. En el seguiente cuadro se resumen los modelos :

Variables explicativas / Variables a explicar	MODELO		
	FACTORES (Cualit.)	COVARIABLES (Cuantit.)	FACTORES Y COVARIABLES (Ambas)
UNA VARIABLE	Análisis de varianza (univariable)	Regresión múltiple (univariable)	Análisis de covarianza (univariable)
VARIAS VARIABLES	Análisis de varianza (multivariable)	Regresión múltiple (multivariable)	Análisis de covarianza (multivariable)

.../...

ALGUNOS RESULTADOS SOBRE LOS METODOS UTILIZADOS EN LOS EJEMPLOS PRESENTADOS

De una manera general, escribiremos el modelo lineal en la forma matricial siguiente :

$$Y = X\beta + \varepsilon$$

donde :

$Y_{(n,1)} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ es el vector de observaciones de una misma variable (caso univariable)

$Y_{(n,q)} = \begin{bmatrix} Y_{11} \cdots Y_{1q} \\ \vdots \quad \vdots \\ Y_{n1} \cdots Y_{nq} \end{bmatrix}$ es la matríz de observaciones de q variables (caso multivariable)

$X_{(n,p)\ (p<n)} = \begin{bmatrix} x_{11} \cdots x_{1p} \\ \vdots \quad \vdots \\ x_{n1} \cdots x_{np} \end{bmatrix}$ es la matríz : - de las variables explicativas en el caso de la regresión
- de las variables indicatrices en el caso del análisis de varianza

En el caso del análisis de covarianza X contiene variables explicativas e indicatrices.

β y ε son respectivamente :

- vector de parámetros y vector de errores en el caso univariable
- matríz de parámetros y matriz de errores en el caso multivariado

En cada ejemplo tratado explicitaremos el vector (matríz) β .

.../...

A - EL MODELO LINEAL : UN ENFOQUE GEOMETRICO

I - LA REGRESION

El modelo :

$$\underset{(n,1)}{Y} = \underset{(n,p)}{X}\ \underset{(p,1)}{\beta} + \underset{(n,1)}{\varepsilon}$$

Hipótesis :

En el caso de la regresión suponemos :

$E(\varepsilon) = 0$ (esperanza del vector error es el vector 0)

Var $(\varepsilon) = \sigma^2 I_n$ (matríz de varianza-covarianza del error)

Errores independientes

Normalidad de los errores.

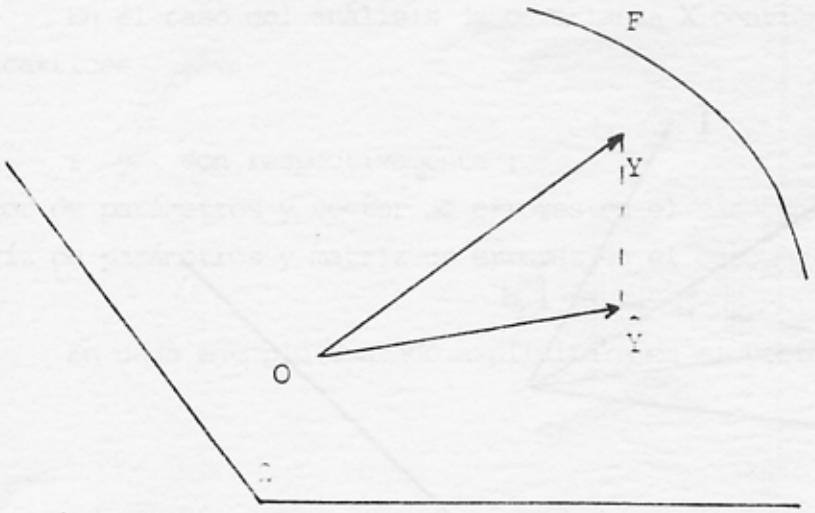
Problemas :

Dentro de los problemas que abordaremos están :

estimación de los parámetros : β

pruebas de hipótesis

Representación geométrica :



Y : vector de observaciones

Ω : s.e.v. de F engendrado por las columnas de X

$\hat{Y}$: proyección ortogonal de Y sobre Ω

F : espacio de variable

.../...

I-1. Caso simple : la regresión lineal

$$Y = X\beta + \varepsilon$$

con

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad ; \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \left[1, x\right] \; ; \; \beta = \begin{bmatrix} b_o \\ b_1 \end{bmatrix} \quad ; \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Esto no es más que la representación matricial del modelo

$$y_i = b_o + b_1 x_i + \varepsilon_i \qquad i = 1, \ldots, n$$

Hipótesis :

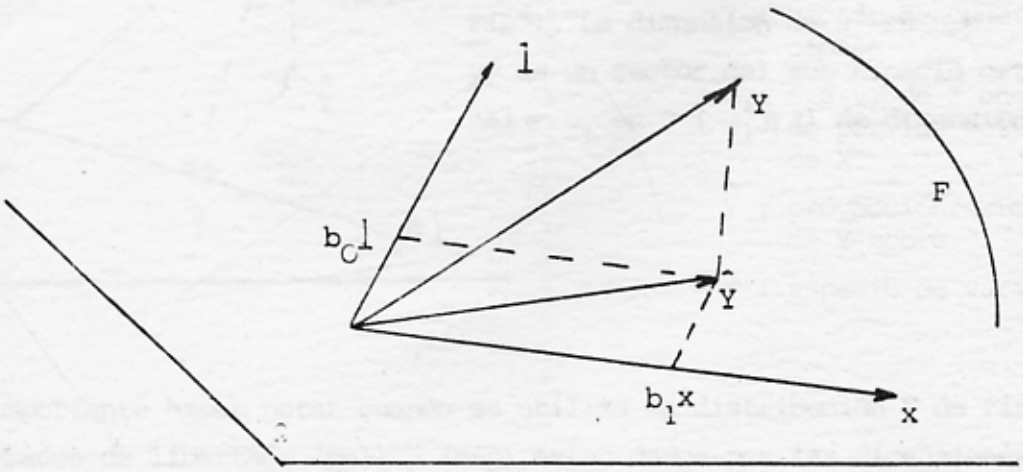
- $E(\varepsilon_i) = 0$
- $\text{Var}\ (\varepsilon_i) = \sigma^2$

i.e. $E(\varepsilon) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$ y $\text{Var}\ (\varepsilon) = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}$

- normalidad de los errores i.e. $\varepsilon_i \sim N(0, \sigma^2)$
- errores independientes i.e. $\text{cov}\ (\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j$

Problemas :

- estimación de los parámetors σ^2, b_o, b_1
- distribuciones de los estimadores $\hat{\sigma}^2, \hat{b}_o, \hat{b}_1$



.../...

I-2. Resultados generales

Estimación :

La estimación de los parametros está dada por :

$\hat{\beta} = ({}^tXX)^{-1}\, {}^tXY$ (écuaciones normales)

$E(\hat{\beta}) = \beta$

$Var(\hat{\beta}) = \sigma^2\, ({}^tXX)^{-1}$

Prueba de hipotesis :

- Prueba general :

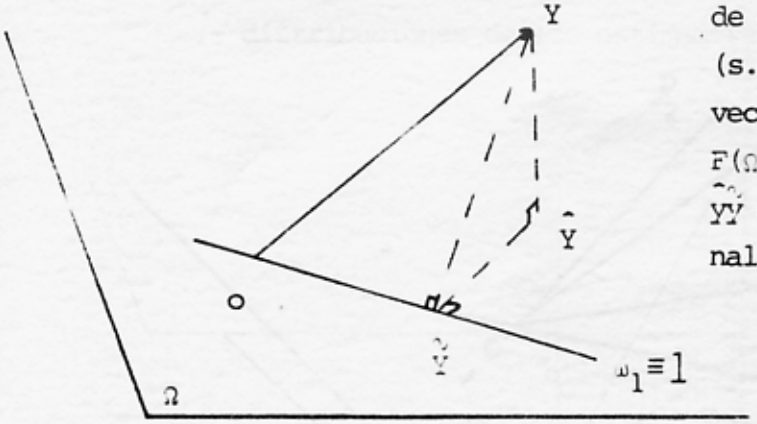
La prueba general pretende saber si todos los coeficientes $b_1,\ldots,b_{p-1}$ de regresión (parámetros) son significativamente diferentes de cero. Esta hipótesis se puede escribir así :

$H_o : E(Y) = b_o \mathbb{1}$

$\bar{H}_o : E(Y) = X\beta$ (hipótesis alternativa)

Para esta prueba se utiliza la estadística :

$$F = \frac{||\hat{\tilde{Y}}||^2/(p-1)}{||\hat{Y}||^2/(n-p)} \sim F_{(p-1,n-p)} \quad \text{bajo } H_o$$



El vector y está en el espacio $F(=R^n)$. Ω es el espacio del modelo a priori de dimension p ; ω_1 sub espacio vectorial (s.e.v.) de Ω de dimension 1. $\hat{Y}$ es un vector del espacio ortogonal a Ω en $F(\Omega^\perp)$. La dimension de $\Omega^\perp$ es n-p. $\hat{\tilde{Y}}$ es un vector del sub espacio ortogonal a ω_1 en Ω ($\omega_1^\perp \cap \Omega$) de dimension p-1.

Es importante hacer notar cuando se utiliza la distribución F de Fisher-Snedecor los grados de libertad (p-1) y (n-p) estan dados por las dimensiones de los subespacios vectoriales $(\omega_1^\perp \cap \Omega)$ e $\Omega^\perp$

.../...

- Prueba de igualdad a cero de un subconjunto de coeficientes de regresión :

Partamos la matriz X del modelo a priori en $X = (X_1, X_2)$

X_1 de dimension (n, p_1)

X_2 de dimension (n, p_2) $\Big\}$ con $p_1 + p_2 = p$

y el vector β en $= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

β_1 de dimension p_1

β_2 de dimension p_2

Probar la hipótesis $H_o : \beta_1 = 0$ contra $\bar{H}_o : \beta_1 \neq 0$

es equivalente a escoger entre los modelos :

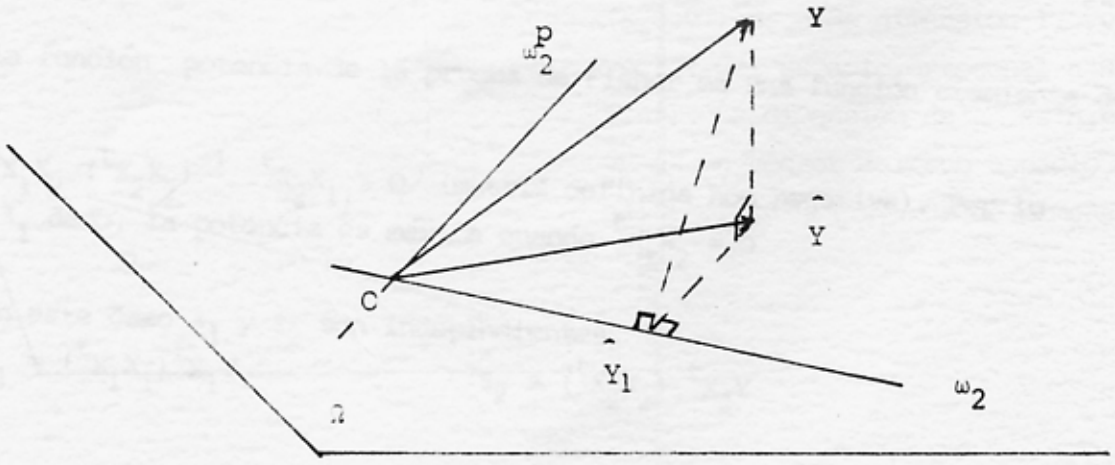
$$H_o : E(Y) = X_2\beta_2$$

$$\bar{H}_o : E(Y) = X_1\beta_1 + X_2\beta_2$$

consideremos el s.e.v. ω_2 engendrado por las columnas de la matriz X_2

De la misma manera que en el caso anterior, se utiliza la estadística :

$$F = \frac{\|\hat{Y} - \hat{Y}_1\|^2 / p_1}{\|Y - \hat{Y}\|^2 / (n-p)} \sim F_{(p_1,\ n-p)} \qquad \text{bajo } H_o$$



con $\omega_2^p = \omega_2^{\perp} \cap \Omega$, dim $\omega_2 = p_2$; dim $\omega_2^p = p - p_2 = p_1$

Así la regla de decisión que se utiliza es la siguiente :

si $F > \delta_\alpha$ se acepta $\bar{H}_o$

y si $F \leq \delta_\alpha$ se acepta H_o

con δ_α tal que $P\{F_{(p_1,\ n-p)} > \delta_\alpha / H_o\} = \alpha$

I-3. Ortogonalidad en regresión

Cuando partimos la matriz X del modelo a priori en $X = (X_1, X_2)$, las ecuaciones normales se escriben :

$$\begin{bmatrix} {}^tX_1X_1 & {}^tX_1X_2 \\ {}^tX_2X_1 & {}^tX_2X_2 \end{bmatrix} \cdot \begin{bmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{bmatrix} = \begin{bmatrix} {}^tX_1Y \\ {}^tX_2Y \end{bmatrix}$$

bajo la hipótesis alternativa $\bar{H}_o : \beta_1 \neq 0$, la estadística F se distribuye como un Fisher decentrada $F'_{(p_1,\ n-p,\ \lambda)}$.

$$\lambda = \frac{1}{\sigma^2}\ {}^t\beta_1 ({}^tX_1X_1 - {}^tX_1X_2\ ({}^tX_2X_2)^{-1}\ {}^tX_2X_1)\ \beta_1$$

es el parámetro de no centralidad.

La función potencia de la prueba de Fisher es una función creciente de λ

${}^tX_1X_2\ ({}^tX_2X_2)^{-1}\ {}^tX_2X_1 \geq 0$ (matriz definida non negativa). Por lo tanto, para β_1 dado, la potencia es máxima cuando ${}^tX_1X_2 = 0$

En este caso β_1 y β_2 son independientes.

$\hat{\beta}_1 = ({}^tX_1X_1)\ {}^tX_1Y$ $\qquad$ $\hat{\beta}_2 = ({}^tX_2X_2)\ {}^tX_2Y$

.../...

Estos no son estimadores sesgados en el modelo

$$E(Y) = X_1\beta_1 + X_2\beta_2$$

Se obtiene la suma de cuadrados explicados por el modelo sumando los cuadrados explicados por X_1 y X_2. Esta descomposición aditiva de la suma de los cuadrados explicados por el modelo corresponde a la ortogonalidad de los s.e.v. ω_1^p y ω_2^p.

$$\text{cov}\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \sigma^2 \begin{bmatrix}({}^tX_1X_1)^{-1} & 0\\ 0 & ({}^tX_2X_2)^{-1}\end{bmatrix}$$

Los dos conjuntos de estimadores son no correlacionados (independencia estadística bajo la hipótesis de normalidad).

De lo anterior, se ve el interés de buscar condiciones de ortogonalidad cuando se planifica una experiencia.

II - ANALISIS DE LA VARIANZA

El modelo, las hipótesis, los problemas y la representación geométrica son los mismos que en el caso de la regresión, pero como veremos en el ejemplo que sigue la matríz X tiene una estructura muy particular.

II-1. Caso simple : diseño experimental a dos factores sin interación y sin repeticiones

Se quiere comparar los efectos de tres tipos de abono sobre el rendimiento del trigo en dos tipos de suelo. En la siguiente tabla tenemos el rendimiento observado en las n=6 parcelas.

		ABONO		
		1	2	3
SUELO	1	Y_{11}	Y_{12}	Y_{13}
	2	Y_{21}	Y_{22}	Y_{23}

- el factor "suelo" toma dos modalidades
- el factor "abono" toma tres modalidades

Si suponemos que los efectos son aditivos, se propone el modelo :

$$Y_{ij} = \mu + \beta_i + \gamma_j + \varepsilon_{ij} \qquad \begin{matrix} i = 1, 2 \\ j = 1,2,3 \end{matrix}$$

con μ : término general

β_i: efecto del suelo i

γ_j: efecto del abono j

$\mu + \beta_i + \gamma_j$ es interpretado como el rendimiento medio que se obtiene en las parcelas bajo la influencia simultánea de las modalidades i y j de los factores suelo y abono.

ε_{ij} : es el término aleatorio

El modelo se puede escribir en la forma matricial así :

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix}$$

es decir : $Y = X\beta + \varepsilon$

Observamos que en la matriz X, la primera columna es una combinación lineal de las otras. Aquí la matriz X est de rango r (=4) con r inferior a p(=6). Lo que hace que la matriz ${}^tXX$ no sea invertible como en el caso de la regresión Entonces estamos obligados a escoger une solución β_o particular, seleccionando una inversa generalizada de ${}^tXX$.

.../...

II-2. Resultados generales

Estimación :

Como lo hemos dicho anteriormente se hace necesario en el análisis de la varianza (lo mismo que en el análisis de covarianza) escoger una inversa generalizada de la matriz ${}^tXX$. Nosotros introduciremos esta noción de la forma siguiente :

Si la matriz A no es invertible se busca una matriz A^- tal que $AA^-A = A$. La matríz A^- se llama inversa generalizada de A. El lector que desee más información sobre el tema puede consultar a : S.R. SEARLE (9).

Así tenemos :

$$\hat{\beta} = ({}^tXX)^- \; {}^tXY$$

$$E(\hat{\beta}) = ({}^tXX)^- \; {}^tXX\beta$$

$$Var(\hat{\beta}) = ({}^tXX)^- \; {}^tXX \; ({}^tXX)^- \; \sigma^2$$

Prueba de hipótesis :

Dentro de las hipótesis que se pueden hacer está la de probar si hay diferencias significativas entre los niveles de un factor. En el ejemplo anterior una prueba de hipótesis puede ser :

$$H_o : \gamma_j = 0 \qquad \forall\, j$$

$$\bar{H}_o : \gamma_j \neq 0 \qquad \text{para algún } j$$

eso es equivalente a escoger entre los modelos

$$H_o : E(y_{ij}) = \mu + \beta_i$$

$$\bar{H}_o : E(y_{ij}) = \mu + \beta_i + \gamma_j$$

De una manera general escribiremos estas hipótesis de la manera siguiente :

$$H_o : E(Y) = X_1\beta$$

$$\bar{H}_o : E(Y) = X\beta$$

donde la matríz X_1 se escoge en forma apropriada.

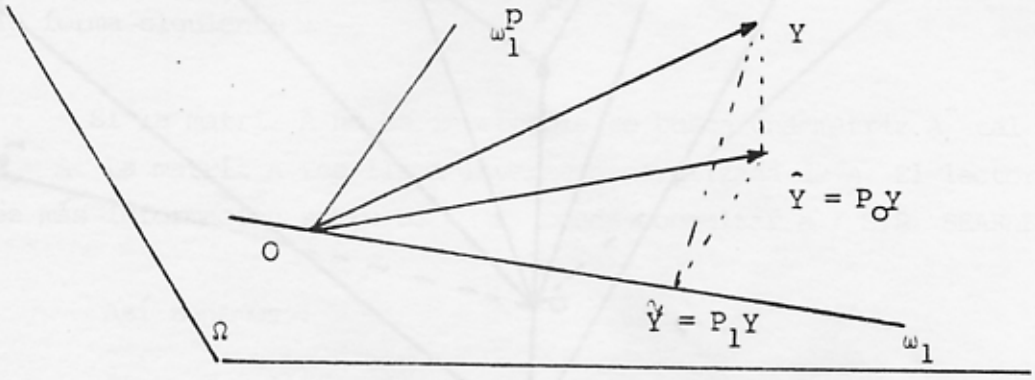
.../...

La estadística utilizada es la siguiente :

$$F = \frac{{}^t Y(P_o - P_1)\, Y \,/\, r_1}{{}^t Y(I - P_o)\, Y/(n-r)} = \frac{||\hat{Y} - \hat{\hat{Y}}||/r_1}{||Y - \hat{Y}||/(n-r)}$$

donde P_o es un proyector ortogonal sobre Ω

P_1 es un proyector ortogonal sobre ω_1



con dim Ω= r (rango de ${}^t XX$), $\omega_1^{\perp} \cap \Omega = \omega_1^p$

dim $\omega_1^p = r_1$, dim $\omega_1 = r - r_1$, dim $\Omega^{\perp} = n - r$

La regla de decisión que se utiliza es :

$F > \zeta_\alpha$ se decide $\bar{H}_o$

$F \leq \zeta_\alpha$ se decide H_o

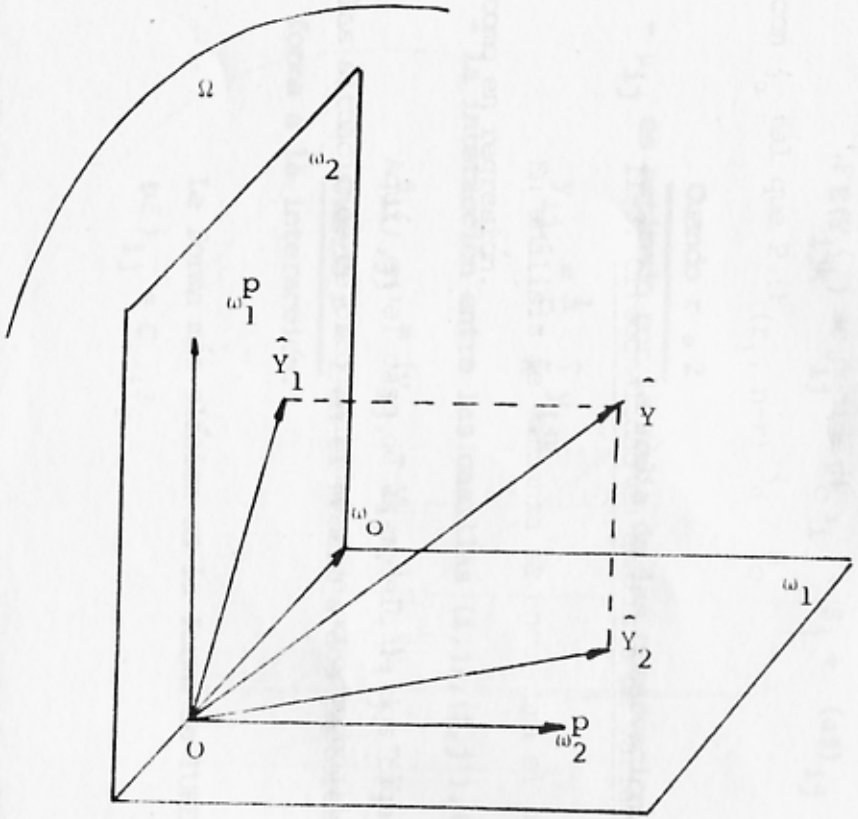
con ζ_α tal que $P\{F_{(r_1,\, n-r)} > \zeta_\alpha / H_o\} = \alpha$

II-3. Ortogonalidad en análisis de varianza

En análisis de varianza se encuentra el problema de no ortogonalidad como en regresión.

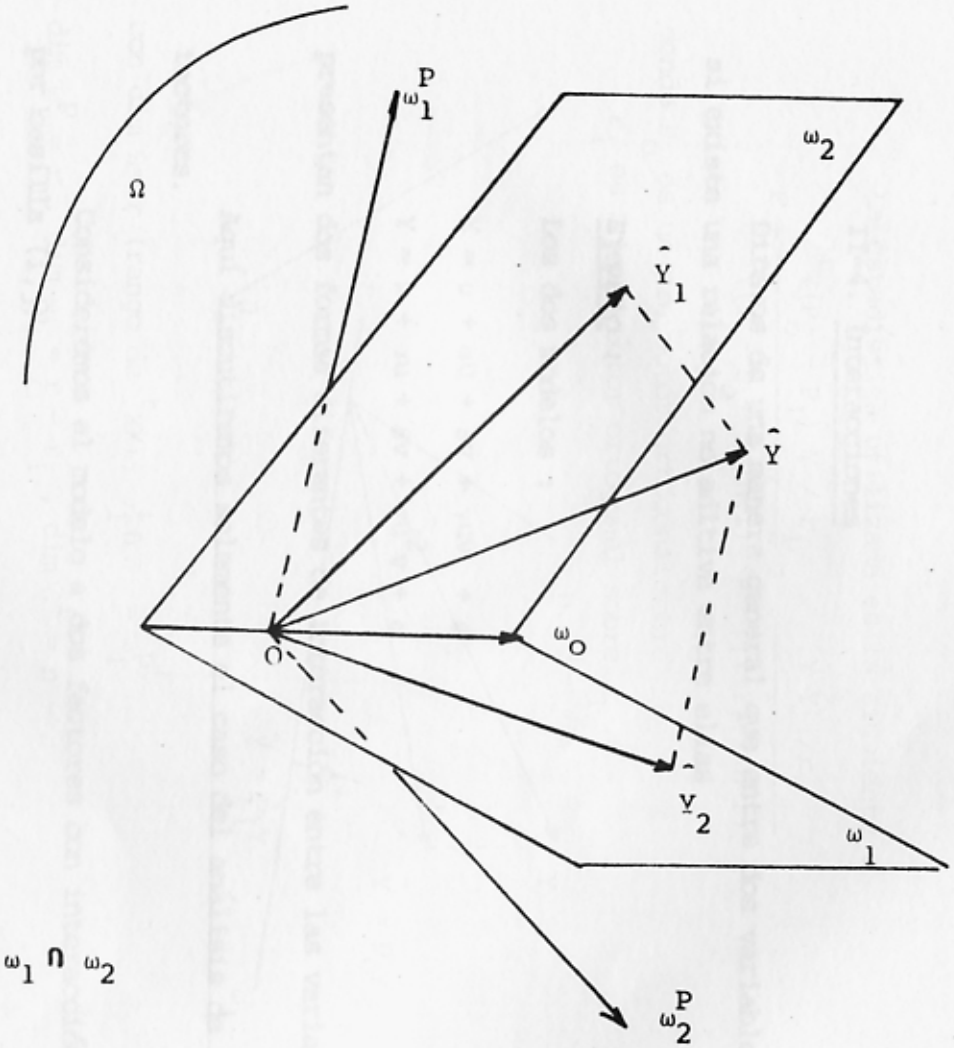
Aquí, en el caso de un modelo de dos factores, vamos a representar las situaciones de ortogonalidad y de no ortogonalidad.

.../...



ORTOGONALIDAD

$$\omega_o = \omega_1 \cap \omega_2$$



NO ORTOGONALIDAD

II-4. Interacciones

Diremos de una manera general que entre dos variables hay interacción si existe una relación no aditiva entre ellas

Ejemplo :

Los dos modelos :

$$Y = \mu + \alpha u + \beta v + \gamma uv + \varepsilon$$

$$Y = \mu + \alpha u + \beta v + \gamma u^2 v + \varepsilon$$

presentan dos formas diferentes de interacción entre las variables y y v.

Aquí discutiremos solamente el caso del análisis de la varianza a dos factores.

Consideremos el modelo a dos factores con interacción y r repeticiones por casilla (i,j).

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \qquad \begin{matrix} i = 1,\ldots, I \\ j = 1,\ldots, J \\ k = 1,\ldots, r \end{matrix}$$

$$E(Y_{ijk}) = \mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Cuando $r \geq 2$

- μ_{ij} es estimado por la media de las observaciones de la casilla (i,j) :

$$Y_{ij.} = \frac{1}{r} \sum_k Y_{ijk}$$

- la interacción entre las casillas (i,j),(i,j'),(i',j),(i',j') es estimado por :

$$\theta_{ii',jj'} = (Y_{ij.} - Y_{ij'.}) - (Y_{i'j.} - Y_{i'j'.})$$

Cuando $r = 1$ en el modelo a dos factores, es necesario dar una forma a la interacción.

La forma más clásica es la forma de TUKEY :

$$(\alpha\beta)_{ij} = C\, \alpha_i \beta_j$$

.../...

Se pueden utilizar otros métodos más sofisticados :

Método de TUKEY generalizada

Método de MANDEL, JOHNSON y GRAYBILL

Todos estos métodos suponen a priori formas de interacción.

Un problema de interpretación de interacción

Un plan de experiencia no es necesariamente adaptado para evidenciar la interacción entre dos factores. Si la casilla (i,j) esta vacía, μ_{ij} no es estimable a no ser que se hagan hipótesis suplementarias sobre μ_{ij}.

Consideremos el plan de experiencia siguiente.

ler factor \ 2do factor	1	2	3
1	2	2	
2		2	2
3	2		2

n = 12 observaciones

No de g.d.l. del ler factor : $3 - 1 = 2$

No de g.d.l. del 2do factor : $3 - 1 = 2$

No de g.d.l. de la interacción : $6 - 3 - 3 + 1 = 1$

No de g.d.l. de la media general : $1 = 1$

6

de donde los g.d.l. del error es : $12 - 6 = 6$

Se puede utilizar el modelo a dos factores con interacción

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

pues se dispone de un g.d.l. para la interacción pero este plan de experiencia es inadecuado para estudiar la interacción $(\alpha\beta)$.

En este caso se puede calcular la suma de los cuadrados de la interacción y del error. Pero la prueba de FISHER utilizando estos valores no se adapta a la hipótesis de ausencia de interacción :

$$Ho : (\mu_{ij} - \mu_{i'j}) - (\mu_{ij'} - \mu_{i'j'}) = 0 \qquad \forall\ i, i', j, j'$$

¿ A cual hipótesis la prueba de FISHER esta adaptada ?

casilla (2,1) vacía : $\theta_{11,22} = (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$ no es estimable

casilla(2,1) y (3,2)vacías : $\theta_{21,32} = (\mu_{21} - \mu_{22}) - (\mu_{31} - \mu_{32})$ no es estimable

casilla (3,2) vacía : $\theta_{22,33} = (\mu_{22} - \mu_{23}) - (\mu_{32} - \mu_{33})$ no es estimable

Pero $\sum\theta = \theta_{11,22} + \theta_{21,32} + \theta_{22,33}$ es estimable

En efecto, $\sum\theta = \mu_{11} - \mu_{12} - \mu_{31} + \mu_{22} - \mu_{23} + \mu_{33}$

y se estime por :

$$\widehat{\sum\theta} = Y_{11.} - Y_{12.} - Y_{31.} + Y_{22.} - Y_{23.} + Y_{33.}$$

bajo la hipótesis $\sum\theta = 0$, $\widehat{\sum\theta}$ se distribuye $N(0, 3\sigma^2)$

$$\frac{\widehat{\sum\theta}}{\sqrt{3\sigma^2}} \sim N(0,1) \; ; \quad \frac{(\widehat{\sum\theta})^2}{3\hat{\sigma}^2} \sim F(1,6)$$

$$F = \frac{(\widehat{\sum\theta})^2}{3} \Big/ \frac{\sum_{ijk} (Y_{ijk} - Y_{ij.})^2}{6}$$ esta adaptada a la prueba $\sum\theta = 0$

En general este estadística se utiliza para hacer la prueba de ausencia de interacción. Pero en el caso que falten datos (casillas vacías) se efectúa en realidad la prueba de que cierta combinación lineal es igual a cero.
- Cuando la prueba es significativa, hay interacción pero no se sabe de que tipo ;
- Cuando la prueba no es significativa, se decide que $\sum\theta = 0$, pero no podemos decir que no hay interacción.

Hemos presentado este caso en el cual hay casillas vacías para sensibilizar los utilizadores de l'análisis de varianza sobre los problemos de interpretación de las interacciones.

.../...

Prueba de hipótesis de ausencia de interacción (representación geométrica)

Se trata de escoger la hipótesis Ho : ausencia de interacción :

$$Ho : \forall\, i,i',j,j' \quad (\mu_{ij} - \mu_{ij'}) - (\mu_{i'j} - \mu_{i'j'}) = 0$$
$$i \neq i', j \neq j'$$

Se trata entonces de escoger entre los modelos :

$$\bar{Ho} : Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad \text{i.e.} \quad Y = X\beta + \varepsilon$$

$$Ho : Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad \text{i.e.} \quad Y = X_1\beta + \varepsilon$$

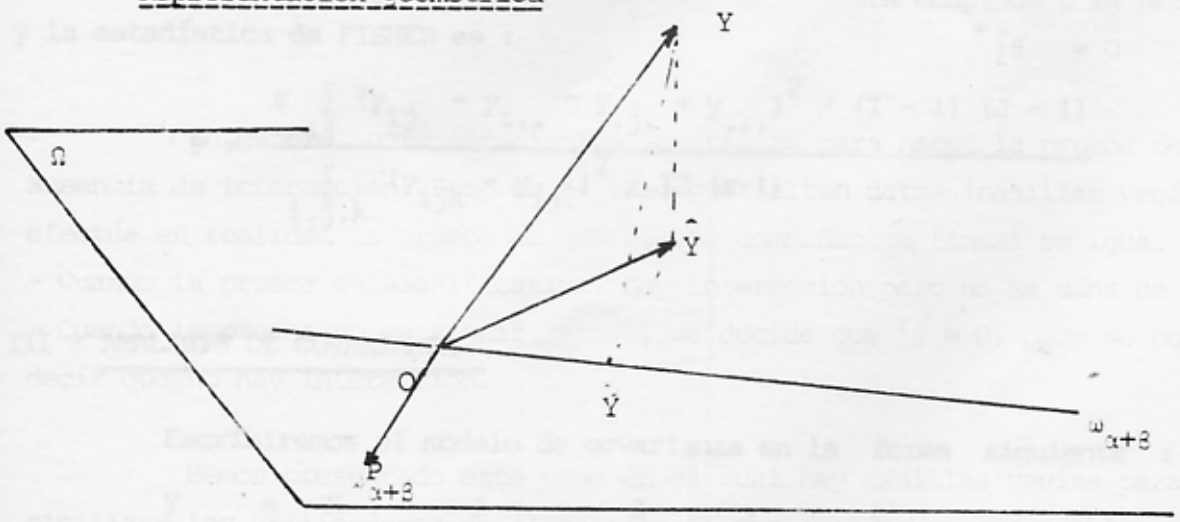
- Sea Ω el s.e.v. de R^n engendrado por las columnas de X. $\dim\Omega = s$ (no. de casillas no vacías)
- Sea $\omega_{\alpha+\beta}$ el s.e.v. engendrado por las columnas de X_1. Se verifica que

$$\dim \omega_{\alpha+\beta} = I + J - 1$$

- $\omega_{\alpha+\beta} \subset \Omega$

Definimos $\omega^P_{\alpha+\beta} = \omega^{\perp}_{\alpha+\beta} \cap \Omega$ $\quad (\dim \omega^P_{\alpha+\beta} = s - I - J + 1)$

Representación geométrica



La estadística utilizada es : $$F = \frac{\hat{Y}\hat{Y}^2 / (S-I-J+1)}{Y\hat{Y}^2 / (n-s)}$$

.../...

Regla de decisión

$$\text{Si : } F > \delta_\alpha \quad \text{se decide } \bar{H}o$$

$$F \leq \delta_\alpha \quad \text{se decide } Ho$$

con δ_α tal que : $P\ (F_{(s-I-J+1,\ n-s)} > \delta_\alpha \ / \ H_o) = \alpha$

Observaciones

$$- \ ||\hat{Y}\bar{Y}||^2 = \sum_i \sum_j \sum_k (y_{ijk} - y_{ij.})^2$$

$$- \ ||\hat{\hat{Y}}\bar{Y}||^2 = ||\hat{Y}||^2 - ||\bar{Y}||^2$$

Se necesita una formulación matricial en el caso general para determinar $\hat{Y}$ y por tanto $||\bar{Y}||^2$ y $||\hat{\hat{Y}}\bar{Y}||^2$.

- En el caso particular de un plan factorial completo equilibrado ($\forall(i,j)$, $n_{ij} = r$) se tiene :

$$||\hat{\hat{Y}}\bar{Y}||^2 = r \sum_{i,j} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2$$

y la estadística de FISHER es :

$$F = \frac{r \sum_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 \ / \ (I-1)\ (J-1)}{\sum_{i,j,k} (y_{ijk} - y_{ij.})^2 \ / \ IJ\ (r-1)}$$

III - ANALISIS DE COVARIANZA

Escribiremos el modelo de covarianza en la forma siguiente :

$$\underset{(n,1)}{Y} = \underset{(n,m)}{X} \ \underset{(m,1)}{\beta} + \underset{(n,k)}{Z} \ \underset{(k,1)}{\gamma} + \underset{(n,1)}{\varepsilon}$$

X : matriz del plan de experiencia

β : vector de parámetros

Z : matriz de las covariables

.../...

γ : vector de coeficientes de regresión

ε : errores independientes, centradas, de varianza σ^2 y normalmente distribuidos

Sea Ω el s.e.v. engendrado por las columnas des $[X,Z]$

ω_1 s.e.v. engendrado por las columnas de X

ω_2 s.e.v. engendrado por las columnas de Z

En general : $\omega_1 \cap \omega_2 = \emptyset$

$$\dim \Omega = \dim \omega_1 + \dim \omega_2$$

$$\dim \Omega^{\perp} = n - \dim \Omega$$

III-1. Resultados generales

Estimación

Estimaciones de los parámetros son soluciones de las ecuaciones normales

$$\begin{bmatrix} t_{XX} & t_{XZ} \\ t_{ZX} & t_{ZZ} \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} t_{XY} \\ t_{ZY} \end{bmatrix}$$

Así tenemos :

$$\hat{\gamma} = ({}^{t}Z Z)^{-1} \; {}^{t}Z(Y - X\hat{\beta})$$

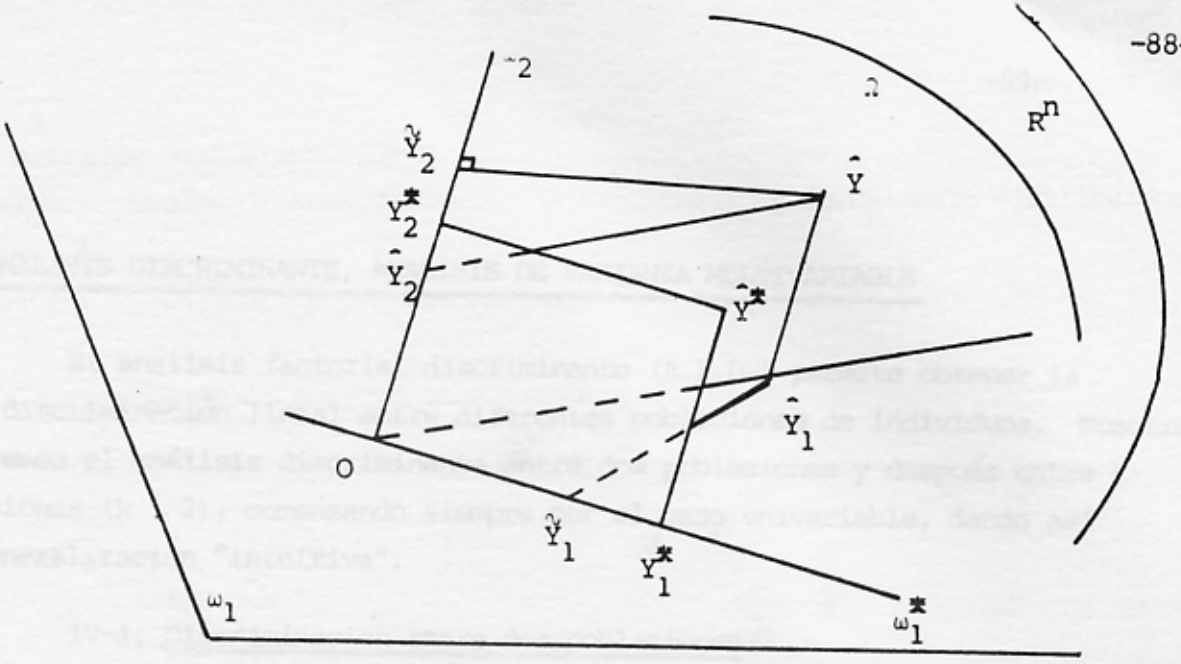
$$\hat{\beta} = \left[{}^{t}X(I - Z({}^{t}ZZ)^{-1} \; {}^{t}Z)X \right]^{-} \cdot \left[{}^{t}X(I - Z({}^{t}ZZ)^{-1} \; {}^{t}Z) \; Y \right]$$

$$\hat{\sigma}2 = \frac{\| Y \|^2}{n - \dim \Omega}$$

Prueba de hipótesis de ausencia de efecto de un factor

Sea ω_1^* el s.e.v. engendrado por las columnas de X. X_* es la matriz del plan de experiencia bajo la hipótesis de ausencia de efecto del factor a pruebar.

.../...



Sea $\hat{Y}$ la proyección de Y sobre Ω

$\hat{Y}^*$ la proyección de Y sobre el s.e.v. engendrado por ω_1^* y ω_2

La estadística utilizado es la siguiente :

$$F = \frac{||\hat{Y}\hat{Y}^*||^2 / \dim (\omega_1^* \oplus \omega_2)^p}{||Y\hat{Y}||^2 / (n - \dim \Omega)}$$

$(\omega_1^* \oplus \omega_2)^p$ es el complementario ortogonal en Ω del s.e.v. $(\omega_1^* \oplus \omega_2)$ engendrado por ω_1^* y ω_2

$$||\hat{Y}\hat{Y}^*||^2 = ||Y\hat{Y}^*||^2 - ||Y\hat{Y}||^2$$

$$||Y\hat{Y}||^2 = \hat{\beta}' t_{XY} + \hat{\gamma}' t_{ZY}$$

$$||Y\hat{Y}^*||^2 = \hat{\beta}_* t_{X_*Y} + \hat{\gamma}_* t_{Z_*Y}$$

$\hat{\beta}_*$ y $\hat{\gamma}_*$ son estimadores de los parámetros del modelo reducido

Prueba de hipótesis de ausencia de efecto de una covariable

La prueba se hace de una misma manera. Aquí,

ω_1 es el s.e.v. engendrado por las columnas de Z

ω_1^* es el s.e.v. engendrado por las columnas de Z_*. Z_* matriz de las covariables que aparecen en el modelo reducido

ω_2 es el s.e.v. engendrado por las columnas de X

.../...

IV - ANALISIS DISCRIMINANTE, ANALISIS DE VARIANZA MULTIVARIABLE

El análisis factorial discriminante (A.F.D.) permite obtener la mejor discriminación lineal entre diferentes poblaciones de individuos. Nosotros trataremos el análisis discriminante entre dos poblaciones y después entre k poblaciones ($k \geq 2$), comenzando siempre por el caso univariable, dando así una generalización "intuitiva".

IV-1. Discriminación entre dos poblaciones

Una variable medida

Sean P_1 y P_2 poblaciones sobre las cuales se ha medido un caracter Y. Se supone que :

$$Y \sim N(\mu_1, \sigma^2) \text{ sobre } P_1$$

$$Y \sim N(\mu_2, \sigma^2) \text{ sobre } P_2$$

Sea n_1 (resp. n_2) el número de observaciones de P_1 (resp. P_2). Notaremos Y_{ij} la variable correspondiente a la medida del individuo j de la población P_i ($1 \leq i \leq 2$, $1 \leq j \leq n_i$).

Estimación

Se estima : μ_i por $\bar{Y}_{i.} = \dfrac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$ $(i = 1,2)$

σ^2 sobre P_1 por : $\hat{\sigma}_1^2 = \dfrac{\sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_{1.})^2}{n_1 - 1}$

σ^2 sobre P_2 por : $\hat{\sigma}_2^2 = \dfrac{\sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_{2.})^2}{n_2 - 1}$

Un estimador sin sesgo de la varianza común a las dos poblaciones es :

$$\hat{\sigma}^2 = \frac{(n_1-1)\,\hat{\sigma}_1^2 + (n_2-1)\,\hat{\sigma}_2^2}{n_1 + n_2 - 2} = \frac{\sum_{j=1}^{n_1} (Y_{1j}-\bar{Y}_{1.})^2 + \sum_{j=1}^{n_2} (Y_{2j}-\bar{Y}_{2.})^2}{n_1 + n_2 - 2}$$

.../...

Para discriminar entre las dos poblaciones se utiliza el criterio de STUDENT :

$$T = \frac{y_{1.} - y_{2.}}{\sqrt{\sigma^2 (\frac{1}{n_1} + \frac{1}{n_2})}}$$

Prueba de hipótesis

Se utiliza, para escoger entre Ho : $\mu_1 = \mu_2$ y $\bar{Ho}$: $\mu_1 \neq \mu_2$ la regla de decisión

$|T| > t \Longrightarrow \bar{Ho}$

$|T| \leq t \Longrightarrow Ho$

Se escoge t de manera a controlar el riesgo de primera clase α

$P \{ |T| > t / H_o \} = \alpha$

o sea

$T^2 > t^2 \Longrightarrow \bar{Ho}$

$T^2 \leq t^2 \Longrightarrow Ho$

Bajo Ho, T es una ley de STUDENT con $n_1 + n_2 - 2$ g.d.l.

Sabemos que :

$$T^2 = \frac{(y_{1.} - y_{2.})^2}{\hat{\sigma}^2 (\frac{1}{n_1} + \frac{1}{n_2})}$$

sigue bajo Ho : $\mu_1 = \mu_2$ la ley de FISHER $F(1, n_1 + n_2 - 2)$

Se puede escribir T^2 bajo la forma :

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2$$

con $D^2 = (y_{1.} - y_{2.}) (\hat{\sigma}^2)^{-1} (y_{1.} - y_{2.})$, distancia entre las medias de las poblaciones P_1 y P_2 con la métrica $\frac{1}{\hat{\sigma}^2}$

Punto de visto geométrico

En realidad, se podrían considerar las poblaciones P_1 y P_2 como niveles de un factor (población) y utilizar el modelo lineal :

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

.../...

Sea en R^n $(n = n_1 + n_2)$ el subespacio vectorial Ω engendrado por las columnas de la matriz :

$$X = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \quad \begin{matrix} \} \, n_1 \text{ - veces} \\ \} \, n_2 \text{ - veces} \end{matrix}$$

Sea $\Omega^\perp$ el espacio ortogonal de Ω en R^n :

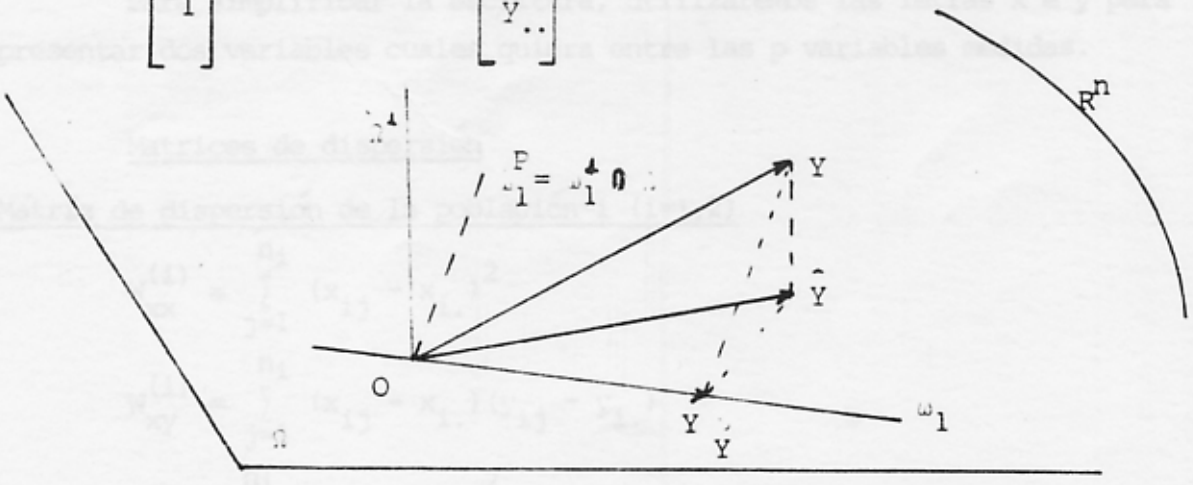
$$\dim \Omega = 2 \ , \ \dim \Omega^\perp = n-2$$

El vector de observaciones Y tiene por proyeccion $\hat{Y}$ sobre :

$$Y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{bmatrix} \qquad \hat{Y} = \begin{bmatrix} y_{1.} \\ \vdots \\ y_{1.} \\ y_{2.} \\ \vdots \\ y_{2.} \end{bmatrix} \quad \begin{matrix} \} \, n_1 \text{ - veces} \\ \} \, n_2 \text{ - veces} \end{matrix}$$

Sea ω_1 el s.e.v. engendrado por 1 ; $\hat{Y}$ la proyección de Y sobre ω_1

$$1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{matrix} n \\ \text{veces} \end{matrix} \qquad Y = \begin{bmatrix} y_{..} \\ \vdots \\ y_{..} \end{bmatrix} \begin{matrix} n \\ \text{veces} \end{matrix} \qquad \text{donde } y_{...} = \frac{1}{n_1 + n_2} \sum_{ij} y_{ij}$$



Utilizando la métrica Euclideana clásica en R^n tenemos :

.../...

$$||\hat{YY}||^2 = \sum_{j=1}^{n_1} (y_{1j} - y_{1.})^2 + \sum_{j=1}^{n_2} (y_{2j} - y_{2.})^2$$

$$||\tilde{YY}||^2 = \sum_{i=1}^{2} n_1 (y_{i.} - y_{..})^2$$

La estadística utilizada bajo Ho es :

$$F(1, n_1 + n_2 - 2) = \frac{\|\tilde{YY}\|^2/1}{\|\hat{YY}\|^2/(n_1+n_2-2)}$$

P variables medidas

Buscamos discriminar entre P_1 y P_2 utilizando las p medidas hechas sobre los $n = n_1 + n_2$ individuos.

Sea y_i (i= 1,2) el vector de p variables

$$y_i = \begin{bmatrix} y_i^{(1)} \\ \vdots \\ y_i^{(p)} \end{bmatrix} \sim N_p(M_i, \Sigma) \qquad \text{donde } M_i = \begin{bmatrix} \mu_i^{(1)} \\ \vdots \\ \mu_i^{(p)} \end{bmatrix}$$

con $y_i^{(j)}$ (resp. $\mu_i^{(j)}$) jésima variable (resp. jésima media teórica) de la población i(i=1,2) y Σ la matriz de varianza-covarianza.

Para simplificar la escritura, utilizaremos las letras x e y para representar dos variables cuales quiera entre las p variables medidas.

Matrices de dispersión

- Matríz de dispersión de la población i (i=1,2)

$$W_{xx}^{(i)} = \sum_{j=1}^{n_i} (x_{ij} - x_{i.})^2$$

$$W_{xy}^{(i)} = \sum_{j=1}^{n_i} (x_{ij} - x_{i.})(y_{ij} - y_{i.})$$

$$W_{yy}^{(i)} = \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2$$

.../...

Se arreglan estos elementos en la matríz $W^{(i)}$ de dimensión (p x p) a la cual están asociados $n_i - 1$ g.d.l.

$$W^{(i)} = \begin{bmatrix} \ddots & & \ddots \\ & \ddots & \vdots \\ W_{xx}^{(i)} & \cdots\cdots & W_{xy}^{(i)} \\ \vdots & & \vdots \\ W_{yx}^{(i)} & \cdots\cdots & W_{yy}^{(i)} \\ \vdots & & \ddots \\ \ddots & & \end{bmatrix}$$

$W^{(i)}$ es la matríz de inercía en R^n de los puntos de P_i, i = 1,2.

- Matríz de dispersión común : INTRA

$$W = W^{(1)} + W^{(2)}$$

i.e.
$$W_{xx} = W_{xx}^{(1)} + W_{xx}^{(2)}$$
$$W_{xy} = W_{xy}^{(1)} + W_{xy}^{(2)}$$
$$W_{yy} = W_{yy}^{(1)} + W_{yy}^{(2)}$$

A la matríz W están asociados $n_1 + n_2 - 2$ g.d.l. , y $\frac{1}{n_1 + n_2 - 2}$ W estima sin sesgo la matríz de covarianza Σ .

- Matríz de dispersión entre poblaciones : INTER

$$B_{xx} = n_1(x_{1.} - x_{..})^2 + n_2(x_{2.} - x_{..})^2$$

$$B_{xy} = n_1(x_{1.} - x_{..})(y_{1.} - y_{..}) + n_2(x_{2.} - x_{..})(y_{2.} - y_{..})$$

$$B_{yy} = n_1(y_{1.} - y_{..})^2 + n_2(y_{2.} - y_{..})^2$$

Se arreglan estos elementos en la matriz B de dimension (p x p) con 1 g.d.l.

.../...

$$B = \begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & B_{xx} & \cdots & B_{xy} & \cdots \\ & \vdots & & \vdots & \\ \cdots & B_{yx} & \cdots & B_{yy} & \cdots \\ \iddots & \vdots & & \ddots & \end{bmatrix}$$

B es la matríz de inercia en R^n de los puntos "media" de las dos poblaciones.

Un cálculo elemental permite demostrar que :

$$\frac{n_1 n_2}{n_1 + n_2} (x_{1.} - x_{2.})(y_{1.} - y_{2.}) = n_1 (x_{1.} - x_{..})(y_{1.} - y_{..}) + n_2 (x_{2.} - x_{..})(y_{2.} - y_{..})$$

de donde :

$$B_{xx} = \frac{n_1 n_2}{n_1 + n_2} (x_{1.} - x_{2.})^2$$

$$B_{xy} = \frac{n_1 n_2}{n_1 + n_2} (y_{1.} - y_{2.})(x_{1.} - x_{2.})$$

$$B_{yy} = \frac{n_1 n_2}{n_1 + n_2} (y_{1.} - y_{2.})^2$$

Se definimos por :

$$d = \begin{bmatrix} \vdots \\ x_{1.} - x_{2.} \\ \vdots \\ y_{1.} - y_{2.} \\ \vdots \end{bmatrix} \qquad \text{entonces} \qquad B = \frac{n_1 n_2}{n_1 + n_2} \; d \; {}^t d$$

- Matríz de dispersión total

$$T = B + W$$

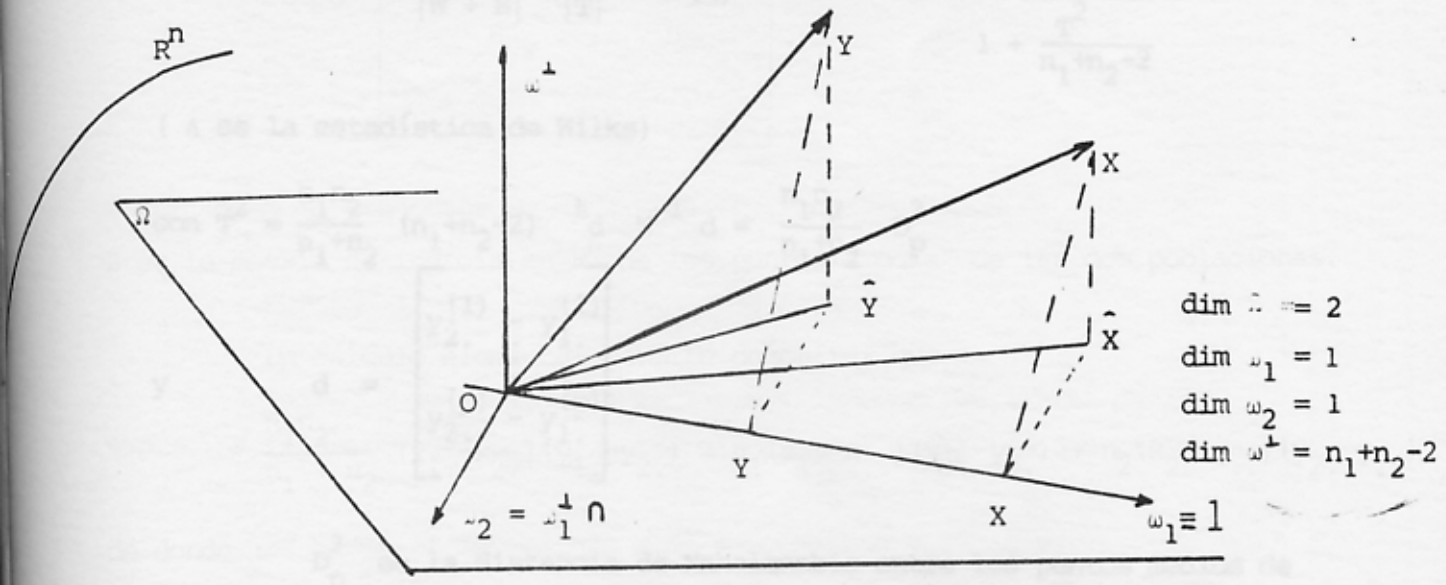
$$T_{xx} = B_{xx} + W_{xx} = \sum_i \sum_j (x_{ij} - x_{..})^2$$

$$T_{xy} = B_{xy} + W_{xy} = \sum_i \sum_j (x_{ij} - x_{..})(y_{ij} - x_{..})$$

$$T_{yy} = B_{yy} + W_{yy} = \sum_i \sum_j (y_{ij} - y_{..})^2$$

T es la matríz de inercia de los puntos en R^n, a la cual están asociados n - 1 g.d.l.

Representación geométrica (métrica Euclideana clásica)



$$W_{xx} = ||\hat{XX}||^2 \qquad B_{xx} = ||\hat{XX}||^2$$
$$W_{yy} = ||\hat{YY}||^2 \quad \text{INTRA} \qquad B_{yy} = ||\hat{YY}||^2 \quad \text{INTER}$$
$$W_{xy} = \langle \hat{XX}, \hat{YY} \rangle \qquad B_{xy} = \langle \hat{XX}, \hat{YY} \rangle$$

$$T_{xx} = ||XX||^2$$
$$T_{yy} = ||YY||^2 \quad \text{TOTAL}$$
$$T_{xy} = \langle XX, YY \rangle$$

Prueba de hipótesis $H_o : M_1 = M_2$ contra $\bar{H}_o : M_1 \neq M_2$

Bajo la hipótesis $H_o : M_1 = M_2$ (de igualdad de distribuciones de las poblaciones P_1 y P_2 sobre los p caracteres medidos), T es la matríz INTRA residual.

.../...

La estadística de la prueba se obtenida por el método del cociente de verosimilitud[1].

Sea λ el cociente de verosimilitud

$$\Lambda = \frac{|W|}{|W+B|} = \frac{|W|}{|T|} = (\lambda)^{\frac{2}{n_1+n_2}} = \frac{1}{1+\frac{T^2}{n_1+n_2-2}}$$

(Λ es la estadística de Wilks)

$$\text{con } T^2 = \frac{n_1 n_2}{n_1+n_2}(n_1+n_2-2)\ t_d\ W^{-1}\ d = \frac{n_1 n_2}{n_1+n_2}\ D_p^2$$

$$\text{y} \qquad d = \begin{bmatrix} y_{2.}^{(1)} - y_{1.}^{(1)} \\ \vdots \\ y_{2.}^{(p)} - y_{1.}^{(p)} \end{bmatrix}$$

D_p^2 es la distancia de Mahalanobis entre los puntos medios de las poblaciones P_1 y P_2 ; la métrica usada es W_c^{-1} con $W_c = \frac{1}{n_1+n_2-2}\ W$.

$\frac{n_1 n_2}{n_1+n_2}\ D_p^2$ es el equivalente multivarable del cuadrado de la estadística de STUDENT ; $T^2 = \frac{n_1 n_2}{n_1+n_2}\ D_p^2$ es la estadística de HOTELLING.

Regla de decisión

$\Lambda_\alpha < \Lambda \leq 1$ se decide Ho : $M_1 = M_2$

$\Lambda \leq \Lambda_\alpha$ se decide $\bar{H}o$

De una manera equivalente se tiene :

$T^2 > \tau_\alpha$ se decide $\bar{H}_o$

$T^2 \leq \tau_\alpha$ se decide H_o

.../...

(1) Cociente de verosimilitud : Sea L la función de verosimilitud y λ el cociente : $\lambda = \frac{\sup_{H_o} L}{\sup L}$

Así se utiliza la regla de decision: \ si $\ell_\alpha < \lambda \leq 1 \Rightarrow Ho$
/ si $\lambda \leq \ell_\alpha \Rightarrow \bar{H}o$

con $\Lambda_\alpha = \dfrac{1}{1 + t_\alpha/(n_1+n_2-2)}$

Se puede mostrar que la estadística

$$\frac{n_1+n_2-p-1}{p} \cdot \frac{1}{n_1+n_2-2} \cdot \frac{n_1 n_2}{n_1+n_2} D_p^2$$

sigue (bajo la hipótesis Ho:$M_1=M_2$) la distribución $F(p, n_1+n_2-p-1)$

Función lineal discriminante de FISHER

Queremos determinar la variable Z combinación lineal de las variables iniciales $Y^{(k)}$ que permite discriminar mejor las dos poblaciones.

Se trata de determinar V de componentes $v_1, \ldots, v_i, \ldots, v_p$ tal que la variable $Z = \sum_1^p v_i Y^{(i)}$ maximize el cociente de la variabilidad INTER y de la variabilidad INTRA.

De esta manera se determina que V es tal que :

$$W.V = d \iff V = W^{-1} d$$

$$t_{Vd} = {}^t d\, W^{-1} d = D_p^2 / (n_1 + n_2 - 2)$$

IV-2. Discriminación entre k poblaciones

Una variable medida

Sean k poblaciones $P_1, P_2, \ldots, P_k$ sobre las cuales se ha medido un caracter Y. Suponemos que $Y \sim N(\_i, \_^2)$ sobre la población P_i.

Sea n_i el número de observaciones de la población P_i.

Notaremos y_{ij} la variable corespondiente a la medida del individuo j de la población i.

.../...

Estimación

Para la población P_i :

Se estima : μ_i por : $y_{i.} = \dfrac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$ $(i=1,\ldots,k)$

σ_i^2 por : $\hat{\sigma}_i^2 = \dfrac{\sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2}{n_i - 1}$

Un estimador sin sesgo de la varianza común a las k poblaciones es :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} (n_i-1)\,\hat{\sigma}_i^2}{\sum_{i=1}^{k} (n_i-1)} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2}{\sum_{i=1}^{k} n_i - k}$$

Para descriminar globalmente las k poblaciones se utiliza el criterio de FISHER :

$$F = \frac{\dfrac{\sum_{i=1}^{k} n_i (y_{i.} - y_{..})^2}{k-1}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2}{\sum_{i=1}^{k} n_i - k}}$$

Prueba de hipótesis

Ho : $\mu_1 = \mu_2 = \ldots = \mu_k$ contra $\bar{H}o$: $\exists$ (i,i') t.q. $\mu_i \neq \mu_{i'}$

Se utiliza la regla de decisión

$F > \zeta_\alpha \longrightarrow \bar{H}o$

$F \leq \zeta_\alpha \longrightarrow Ho$

con $P\ (F(k-1, \sum n_i - k) > \zeta_\alpha \,/\, H_o) = \alpha$

Representación geométrica

En este caso el s.e.v. $\ldots$ es engendrado por las vectores de la matriz :

.../...

$$X_{(n,k)} = \begin{bmatrix} 1 & 0.....0 \\ \vdots & \vdots \\ 1 & 0.....0 \\ 0 & 1.....0 \\ \vdots & \vdots \\ & 1.....0 \\ & 0.....0 \\ & \vdots \\ & 0.....1 \\ & \vdots \\ 0 & 0.....1 \end{bmatrix} \begin{matrix} \} n_1 \\ \\ \} n_2 \\ \\ \\ \\ \} n_k \end{matrix}$$

con $n = \sum_{i=1}^{k} n_i$

La proyección de Y sobre ω es $\hat{Y}$:

$$\hat{Y} = \begin{bmatrix} y_{1.} \\ \vdots \\ y_{1.} \\ \vdots \\ y_{k.} \\ \vdots \\ y_{k.} \end{bmatrix} \begin{matrix} \} n_1 \\ \\ \\ \} n_K \end{matrix}$$

Sea ω_1 el s.e. engendrado por $\mathbf{1}$; la proyección de Y sobre ω_1 es $\tilde{Y}$

$$\tilde{Y} = \begin{bmatrix} y_{..} \\ \vdots \\ y_{..} \end{bmatrix} \} n$$

Si utilizamos la métrica Euclideana clásica

$$\|\hat{YY}\|^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2$$

$$\|\tilde{YY}\|^2 = \sum_{i=1}^{k} n_i (y_{i.} - y_{..})^2$$

$$F = \frac{\|\hat{YY}\|^2 / (k-1)}{\|\hat{YY}\|^2 / (n-k)}$$

Se hacen ulteriormente comparaciones de medias dos a dos para discriminar las poblaciones, cuando Ho es rechazada.

P variables medidas

Buscamos ahora discriminar k poblaciones utilizando medidas de p variables hechas sobre $n = \sum_{i=1}^{k} n_i$ individuos.

Sea y_i (i=1,....,k) el vector de p variables en la poblacion i

$$y_i \sim N_p (M_i, \Sigma) \quad (i=1,...,k)$$

Como anteriormente se utiliza las letras x, y para representar dos variables cuales quiera.

Matrices de dispersión

- Matríz de dispersión de la población i (i=1,...,k)

La matriz es igual al caso de k = 2

- Matríz de dispersión común : INTRA

$$W = W^{(1)} + W^{(2)} + \ldots + W^{(k)}$$

A esta matríz están asociados $\sum_i (n_i - 1) = n - k$ g.d.l.

- Matríz de dispersion entre poblaciones : INTER

$$B_{xx} = \sum_{i=1}^{k} n_i (x_{i.} - x_{..})^2$$

$$B_{xy} = \sum_{i=1}^{k} n_i (x_{i.} - x_{..}) (y_{i.} - y_{..})$$

$$B_{yy} = \sum_{i=1}^{k} n_i (y_{i.} - y_{..})^2$$

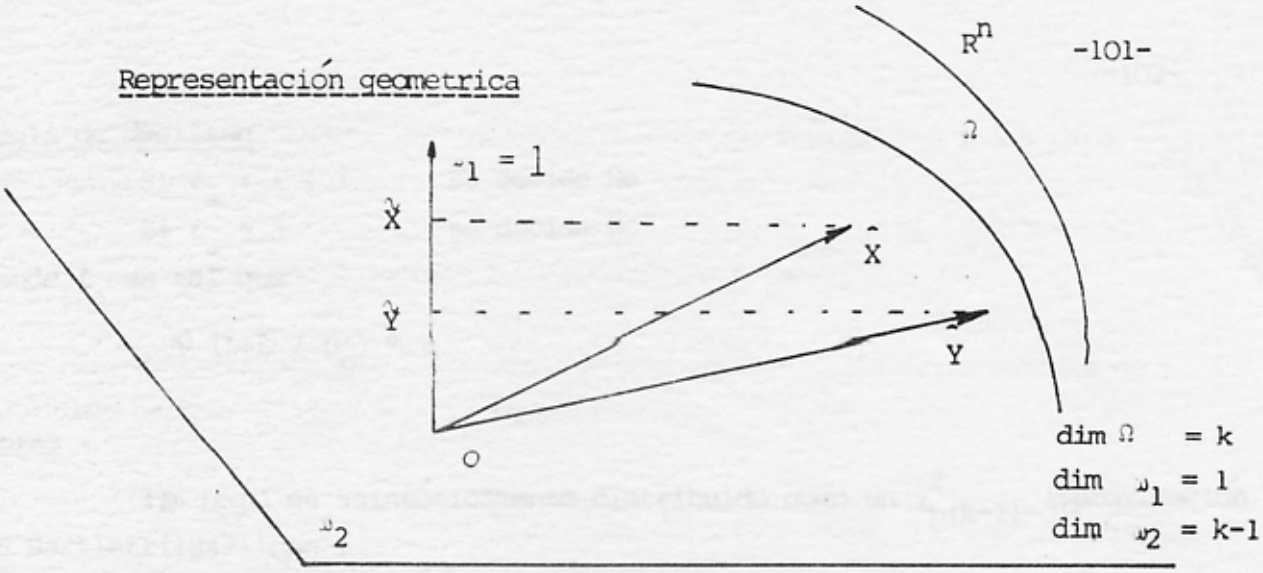
Se arreglan los términos en la matriz B la cual tiene k-1 g.d.l.

$$B = \begin{bmatrix} \ddots & & \vdots & \ddots \\ \cdots & B_{xx} & \cdots & B_{xy} \\ & \vdots & & \\ \cdots & B_{yx} & \cdots & B_{yy} \\ \ddots & & \vdots & \ddots \end{bmatrix}$$

B es la matriz de inercia en R^n de los k puntos "media" de las k poblaciones

.../...

Representación geometrica



$$B_{xx} = \|\hat{X}\hat{X}\|^2$$

$$B_{YY} = \|\hat{Y}\hat{Y}\|^2$$

$$B_{xy} = \langle \hat{X}\hat{X}, \hat{Y}\hat{Y} \rangle$$

Si $p \geq k-1$; $\hat{X}\hat{X}$, $\hat{Y}\hat{Y}$,... engendran "en general" ω_2 de dimension k-1

Si $p < k-1$; $\hat{X}\hat{X}$, $\hat{Y}\hat{Y}$,... engendran "en general" un s.e.v. de ω_2 de dimensión p. i.e.

rango(B) $\leq$ k-1 si $p \geq k-1$

rango(B) $\leq$ p si $p < k-1$

Nota :

La matríz de dispersión total T = B + W es la matríz de productos escalares de las variables $\hat{X}\hat{X}$, $\hat{Y}\hat{Y}$ que se "desplazan" en $\omega_1^{\perp}$ de dimensión n-1.

Prueba de hipótesis Ho : $M_1 = M_2 = \ldots = M_k$ contra $\bar{H}o$: $\exists (i,i') : M_i \neq M_{i'}$

Se trata de decidir entre Ho : las poblaciones $P_1, \ldots, P_k$ tienen la misma distribución y $\bar{H}o$: dos poblaciones al menos son diferentes.

El criterio utilizado es el cociente de verosimilitud λ :

$$\lambda = (\lambda)^{\frac{2}{n}} = \frac{|W|}{|B+W|} = \frac{|W|}{|T|}$$

.../...

Regla de decision

Si $\ell_\alpha < \Lambda < 1$ se decide Ho

Si $\ell_\alpha \geq \Lambda$ se decide $\bar{H}o$

donde ℓ_α es tal que

$$P(\Lambda > \ell_\alpha / H_o) = \alpha$$

Notas :

1) $m \, Log\Lambda$ es asintóticamente distribuido como un $\chi^2_{p(k-1)}$ (aproximación de Bartlett(1947)) con :

$$m = n - 1 - \frac{p+k}{2}$$

Para una informacion más completa ver Rao (5).

2) $\Lambda = \dfrac{|W|}{|T|} = \theta_1 \ldots \theta_r$ con θ_i los valores proprios de $T^{-1}W$

La función discriminante de FISHER : otro criterio de escogencia entre Ho y $\bar{H}o$:

Como para el caso de dos poblaciones se busca la combinación lineal $Z = \sum_1^p v_i Y^{(i)}$ que discrimine mejor las k poblaciones en el sentido del cociente INTER, INTRA.

Variabilidad INTER

$$\sum_1^k n_i (Z_{i.} - Z_{..})^2 = t_{VBV}$$

Variabilidad INTRA

$$\sum_i \sum_j (Z_{ij} - Z_{i.})^2 = t_{VWV}$$

El cociente $\phi = \dfrac{t_{VBV}}{t_{VWV}}$ alcanza un máximo para V_1 vector propio de $W^{-1}B$ correspondiendo al valor propio más grande de θ_1. Así $\phi = \theta_1$.

.../...

El conocimiento de la ley de λ_1 bajo Ho permite construir una regla de decisión adaptada a la escogencia entre Ho y $\bar{Ho}$:

$$\lambda_1 > \ell_\alpha \longrightarrow \bar{Ho}$$

$$\lambda_1 \leq \ell_\alpha \longrightarrow Ho$$

El valor propio más grande es utilizado aquí como criterio de escogencia entre la hipótesis Ho y la hipótesis $\bar{Ho}$.

El vector propio V_1 correspondiente al valor propio más grande se llama el primer eje discriminante.

Se notará $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$ los valores propios de $W^{-1} B$ en orden decreciente y $V_1, \ldots, V_r$ los vectores propios correspondientes. Si los valores $\lambda_2, \ldots, \lambda_s$ son grandes ($s \leq r$) las combinaciones lineales asociadas a $V_2, \ldots, V_s$ son discriminantes.

La hipótesis Ho es demasiado global, así preferimos decidir de la dimensión del espacio que incluye los puntos medios de las k poblaciones, tomando en cuenta la variabilidad INTRA.

Los valores y vectores propios de $W^{-1} B$ son los mismos que los de $T^{-1}B$. Con respecto a la teoría del A.C.P. presentada en el parágrafo B de la parte uno se ve que $V_1, \ldots, V_r$ son formas lineales principales de la nube de puntos medios de las k poblaciones :

Aquí B y T^{-1} son respectivamente los matrices V y M de la teoría del A.C.P. Es decir que el espacio de las unidades estadísticas se provee de la métrica de Mahalanobis.

.../...

Espacio discriminante

Se pretende escoger, con la ayuda de una regla de decisión múltiple, entre las (r+1) hipótesis H_i : dim E = i (i = 0,1,...,r) con E espacio de representación de los puntos "medios".

Daremos dos reglas de decisión adaptadas al problema de decisión múltiple :

Primera regla

Se decide H_i : dim E = i si $\lambda_1 \geq \lambda_2 \geq \ldots > \lambda_i > \ell_\alpha > \lambda_{i+1}$

con ℓ_α tal que $P(\hat{\lambda}_1 > \ell_\alpha / H_0) = \alpha$

Secundo regla

Para escoger entre $H_0, H_1, \ldots, H_r$ se procede así :

- se calcula $\chi^2_0, \chi^2_1, \ldots, \chi^2_{r-1}$ (O estadísticas de Bartllet correspondientes a $B_0, B_1, \ldots, B_{r-1}$).

χ^2	B	g.d.l.
$\chi^2_0 = \lambda_1 + \lambda_2 + \ldots + \lambda_r$	$-(n - \frac{p+k}{2})\{\log\theta_1 + \ldots + \log\theta_r\}$	$p(k-1)$
$\vdots$	$\vdots$	$\vdots$
$\chi^2_j = \lambda_{j+1} + \ldots + \lambda_r$	$-(n - \frac{p+k}{2})\{\log\theta_{j+1} + \ldots + \log\theta_r\}$	$(p-j)(k-j-1)$
$\vdots$	$\vdots$	$\vdots$
$\chi^2_{r-1} = \lambda_r$	$-(n - \frac{p+k}{2})\log\theta_r$	$(p-r-1)(k-r)$

con $(1 + \frac{\lambda_i}{n}) = \frac{1}{\theta_i}$ y $n = \sum n_i - k$

- decidir H_j : dim E = j si

χ^2_j, χ^2_{j+1} .... (resp. $B_j, B_{j+1}\ldots$) son significativamente pequeños y si

$\chi^2_{j-1}, \ldots, \chi^2_0$ (resp. $B_{j-1}, \ldots, B_0$) son significativamente grandes .

.../...

B - EJEMPLOS

PRIMER EJEMPLO : ANALISIS DE LA ALTURA DE LA PRODREDUMBRE EN FUNCION DE CARACTERISTICAS DE LA MUESTRA TOMADA DE TRONCOS DE ARBOLES A 1.30 M DE ALTURA (1).

Altura de la podredumbre en función de características de la muestra tomada de troncos de arboles a 1.30 m de altura.

En este experimento se trata de preveer la altura de la podredumbre en función de :

- diámetro del arbol a 1.3 m
- diámetro de la podredumbre a 1.3 m
- el estado de la podredumbre codificado por 0, 1, 2, 3 según la característica de la gravedad del ataque, apreciada en 56 muestras de troncos.

Para estudiar este problema hemos hecho primeramente estadísticas elementales y después un estudio de regresión.

Estadísticas elementales

No. de la variable	Nombre de la variable	Medias	Desviaciones estandard
1	Ø podredumbre	$0,1196 \times 10^2$	$0,815 \times 10^1$
2	Ø árbol	$0,2859 \times 10^2$	$0,708 \times 10^1$
3	Estado de la podredumbre	$0,1910 \times 10^1$	$0,106 \times 10^1$
4	Altura podredumbre	$0,2995 \times 10^3$	

MATRIZ DE CORRELACIÓN

	1	2	3	4
1	1			
2	.321	1		
3	.656	-.91	1	
4	.818	.144	.734	1

Hemos eligido un programa de regresión múltiple ascendente. La idea de este método es efectuar regresiones : primero tomando solo la variable más correlacionada con la variable a explicar luego con la segunda variable que junto con la primera tiene el coeficiente de correlación múltiple mayor y así succesivamente. Esto se hace con el objeto de que eliminando ciertas variables explicativas se puede mejorar considerablemente la estabilidad del modelo.

Regresión multiple ascendente

Primer nivel

Introducción de la variable 1 (Ø podredumbre) ; es la variable más correlacionada con la variable que se quiere explicar (correlación 0,818).

El resultado del análisis nos da que la altura de la podredumbre :

$$\text{Altura podredumbre} = 0{,}8378 \times 10^2 + 0{,}1218 \times 10^2 \times \text{Ø podredumbre}$$

- desviación standard residual = $0{,}7044 \times 10^2$
- R (correlación entre Y y $\hat{Y}$) = 0,818
- prueba F global : $F = 109{,}23^{**}$ con (1,54) g.d.l.
  ($t = \sqrt{F} = 10{,}45^{**}$)

Cuando el valor F o t aparece con dos asteriscos (**) o un asterisco (*) quiere decir que el coeficiente es significativo al 1% y 5% respectivamente. En caso contrario no es significativo.

Segundo nivel

Introducción de la variable 3 (estado de la podredumbre) ; es la variable explicativa que junto con la variable 1 tiene el coeficiente de correlacion múltiple más grande con la variable a explicar.

$$\text{Altura podredumbre} = 0{,}4899 \times 10^2 + 0{,}8795 \times 10^1 \times \text{Ø podredumbre} + 0{,}3939 \times 10^2 \times \text{Estado podredumbre}$$

- Desviación estandard residual = $0{,}6335 \times 10^2$
- R = 0,858
- $F = 74{,}41^{**}$ con(2,53) g.d.l.
- t de student para los coeficientes de la podredumbre
  $t = 6{,}33^{**}$

t de student para los coeficientes del Estado de la podredumbre
$t = 3,71^{**}$

Tercer nivel

Introducción de la variable 2 (Ø arbol)

$$\text{Altura podredumbre} = 0,5825 \times 10^{2} + 0,8986 \times 10^{1} \times \text{Ø podredumbre} + 0,3824 \times \text{Estado podredumbre} - 0,3262 \times \text{Ø arbol}$$

- Desviación standard residual = $0,6392 \times 10^{2}$
- R = 0,858
- F = $48,74^{**}$ con (3,52) g.d.l.
- t de student (podredumbre) $t = 5.52^{**}$
- t de student (Estado podredumbre) = 3.23^{**}
- t de student (Ø arbol) = - 0,23

CONCLUSIONES :

A través de la regresión ascendente observamos :

- el crecimiento de R (coeficiente de correlación múltiple) conforme se introducen las variables,
- la "estabilización" de los diferentes coeficientes de regresión conforme se introducen las variables 1, 2, 3.
- la disminución de las desviaciones standard residuales (con excepción del nivel tercera)
- la prueba de student en el nivel tercero nos dice que el coeficiente de regresión de la variable 2 (Ø arbol) no es significativo.

En conclusión la ecuación de regresión que de debe usar es la del nivel segundo.

———— + ————

.../...

SEGUNDO EJEMPLO : ANALISIS DEL EFECTO DE UN DESINFECTANTE SOBRE LOMBRICES DE TIERRA EN UN DISENO EXPERIMENTAL EN BLOQUES (2)

En el experimento se trata de saber el efecto de un desinfectante sobre las lombrices de tierra en una parcela dada.De esta manera consideramos :

la variable que se desea explicar y : número de lombrices de tierra por cada 400 gr. de suelo (después del corte) en une parcela dada.

PLAN DE EXPERIENCIA : En bloques "al azar". En el experimento tenemos :

- 4 bloques
- 12 parcelas por bloque ; cada nivel del tratamiento es repetido 1 vez por bloque salvo el nivel "testimonio" que se repite 4 veces.

Los niveles del tratamiento desinfectante son :

1	0 (testigo)
2	1 CN (dosis simple)
3	2 CN (dosis doble)
4	1 CS
5	2 CS
6	1 CM
7	2 CM
8	1 CK
9	2 CK

CN, CS, CM, CK son productos que se aplican en dosis simples y en dosis dobles.

- existe una covariable X = número de lombrices de tierra en 400 g de suelo en la primavera en una parcela dada.

.../...

	0 (1)	2CK	1CN	1CM	2CM	2CS	2CK	0
	269	283	252	212	95	127	80	134
	466	280	398	386	199	166	142	590
	1CS	0	0	2CM	1CK	1CN	1CM	0
BLOQUE 1	138	100	197	263	107	89	41	74
	194	219	421	379	236	332	176	137
	2CS	1CK	0	2CN	0	0	2CN	1CS
	282	230	216	145	88	25	42	62
	372	256	708	304	356	212	308	221
	1CK	0	1CS	2CK	2CK	0	1CK	1CM
	124	211	194	222	193	209	109	153
	268	505	433	408	292	352	132	454
	0	2CN	2CS	1CN	0	2CN	2CS	0
BLOQUE 3	102	193	128	42	29	9	17	19
	363	561	311	222	254	92	28	106
	2CM	0	1CM	0	1CS	1CN	0	2CM
	162	191	107	67	23	19	44	48
	365	563	415	338	80	114	268	298

BLOQUE 2

BLOQUE 4

MODELO 1 :

$$Y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}$$

Y_{ijk} = número de lombrices de tierra observadas en la k-ésima parcela que están localizadas en el bloque j y reciben el tratamiento i.

μ = media general

τ_i = efecto del nivel i del factor "tratamiento" $i = 1,2,\ldots,9$

β_j = efecto del bloque j $j = 1,\ldots, 4$

ε_{ijk} = efecto residual

Se supone que los efectos "tratamientos" y "bloque" son aditivos y que las observaciones son independientes. .../...

(1) Las letras indican el nivel del factor tratamiento. La primera cifra corresponde a "Y" número de lombrices de tierra después del corte. La segunda cifra corresponde a "X" número de lombrices de tierra en primavera.

MODELO 2 :

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma(z_{ijk} - \bar{z}) + \varepsilon_{ijk}$$

Aquí introducimos la covariable centrada $z_{ijk} - \bar{z}$.

ANÁLISIS DEL MODELO 1

Se trata sucesivamente :

- las ecuaciones normales
- el cuadro de análisis de varianza
- estimación de los parámetros
- une descomposición de las sumas de cuadrados ajustados de los tratamientos.

Las ecuaciones normales

μ	β_1	β_2	β_3	β_4	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	SUMAS S	Medias (1) Marginales
48	12	12	12	12	16	4	4	4	4	4	4	4	4	14 680	305.83
	12	0	0	0	4	1	1	1	1	1	1	1	1	4 383	365.25
		12	0	0	4	1	1	1	1	1	1	1	1	3 075	256.25
			12	0	4	1	1	1	1	1	1	1	1	4 752	396.00
				12	4	1	1	1	1	1	1	1	1	2 470	205.83
					16	0	0	0	0	0	0	0	0	5 858	366.13
						4	0	0	0	0	0	0	0	1 066	266.50
							4	0	0	0	0	0	0	1 265	316.25
								4	0	0	0	0	0	928	232.00
									4	0	0	0	0	877	219.25
										4	0	0	0	1 431	357.75
											4	0	0	1 241	310.25
												4	0	892	223.00
													4	1 122	280.50

t_{XX} (modelo 1) t_{XY} (modelo 1)

. t_{XX} es una matriz simétrica y hemos representado solamente la parte superior.

.../...

(1) Cociente de S por el elemento diagonal de la fila correspondiente de t_{XX}.

TABLA DEL ANÁLISIS DE VARIANZA

Fuente de variación	g.d.l.	suma de cuadrados (S.C.)	cuadrados medios (C.M.)	F
TRATAMIENTO	8	157 448	18 681	1.30
BLOQUE	3	289 427	96 476	6.38 **
ERROR	36	544 690	15 130	
TOTAL	47	991 565		

Se obtiene, por ejemplo, fácilmente la suma de cuadrados ajustado del factor "tratamiento" por el cálculo siguiente (ver ecuaciones normales).

$$\frac{(5858)^2}{16} + \frac{(1066)^2}{4} + \ldots + \frac{(1122)^2}{4} - \frac{(14680)^2}{48} = 157448$$

Dado que la prueba de hipótesis F "tratamiento" no es significativa, no es necesario ir más allá con el análisis del modelo 1 . Sin embargo, nosotros seguiremos para preparar el análisis del modelo 2 .

ESTIMACIÓN DE LOS PARAMETROS

Notemos que la matriz ${}^tXX$ es singular. Por lo tanto hay que introducir condiciones suplementarias para dar un sentido a los parámetros[1] .

Para estimar τ_i por $Y_{i..} - Y_{...}$

β_j por $Y_{.j.} - Y_{...}$

μ por $Y_{...}$

se utiliza las condiciones suplementarias :

$$\begin{cases} 4\tau_1 + \tau_2 + \ldots + \tau_j = 0 \\ \beta_1 + \beta_2 + \ldots + \beta_4 = 0 \end{cases}$$

Las estimaciones de los niveles del tratamiento son : .../...

(1) Escoger condiciones suplementarias es equivalente escoger una inversa generalizada de ${}^tXX$ (cf. § II-1 p.79).

Testigo	τ_1	+ 60.30
1 CN	τ_2	- 39.33
2 CN	τ_3	+ 10.42
1 CS	τ_4	- 73.83
2 CS	τ_5	- 86.58
1 CM	τ_6	51.92
2 CM	τ_7	4.42
1 CK	τ_8	- 82.83
2 CK	τ_9	- 25.33

Una descomposición de la S.C. ajustada de los tratamientos :

Consideremos las hipótesis lineales definidas por los métodos :

Testigo	τ_1	- 4	+ 4						
1 CN	τ_2	0	- 2	- 2		- 2	+ 1		+ 1
2 CN	τ_3	+ 1	+ 1	+ 1		+ 1	+ 2		+ 2
1 CS	τ_4	0	- 2	+ 2		- 2	- 1		+ 1
2 CS	τ_5	+ 1	+ 1	- 1		+ 1	- 2		+ 2
1 CM	τ_6	0	- 2		- 2	+ 2		+ 1	- 1
2 CM	τ_7	+ 1	+ 1		+ 1	- 1		+ 2	- 2
1 CK	τ_8	0	- 2		+ 2	+ 2		- 1	- 1
2 CK	τ_9	+ 1	+ 1		- 1	- 1		- 2	- 2
		H_1	H_2	H_3			H_4		

$H_1 \equiv \tau_3 + \tau_5 + \tau_7 + \tau_9 - 4\tau_1 = 0$

(1 g.d.l.) (comparación de la dosis doble del desinfectante con el testigo)

$H_2 \equiv \tau_3 + \tau_5 + \tau_7 + \tau_9 - 2(\tau_2 + \tau_4 + \tau_6 + \tau_8) + 4\tau_1 = 0$

(1 g.d.l.) (linealidad entre los dosis dobles y testigo con la dosis simple).

$$H_3 \equiv \tau_3 - 2\tau_2 = \tau_5 - 2\tau_4$$
$$\tau_7 - 2\tau_6 = \tau_9 - 2\tau_3$$
$$\tau_3 - 2\tau_2 + \tau_5 - 2\tau_4 = (\tau_7 - 2\tau_6 + \tau_9 - 2\tau_8)$$

(3 g.d.l.) misma desviación entre los desinfectantes (CN, CS), (CM, CK) y ((CN,CS),(CM,CK))

$$H_4 \equiv 2\tau_3 + \tau_2 = 2\tau_5 + \tau_4$$
$$2\tau_7 + \tau_6 = 2\tau_9 + \tau_8$$
$$2(\tau_3 + \tau_5) + \tau_2 + \tau_4 = 2(\tau_7 + \tau_9) + \tau_6 + \tau_8$$

(3 g.d.l.) igualdad entre los desinfectantes (CN,CS, (CM,CK) y((CN,CS),(CM,CK))

En estadística una muestra de n observaciones posee n grados de libertad. Pero a partir del momento en que k relaciones independientes existen entre las observaciones el número de g.d.l. es n-k.

Asi por ejemplo la hipótesis H_1 comporta una relación lineal entre las niveles de dosis doble y el testigo y por lo tanto tiene un grado de libertad :

$$\tau_3 + \tau_5 + \tau_7 + \tau_9 - 4\tau_1 = 0$$

Esta hipótesis la escribiremos en forma matricial come sigue :

$$\begin{bmatrix} 0 & 0 & 0 & 0 & -4 & 0 & 1 & 0 & 1 & 0 \\ & & & & & & & & & \\ & & & & 0 & & & & & \end{bmatrix} \begin{bmatrix} u \\ \tau_1 \\ \vdots \\ \tau_4 \\ \tau_1 \\ \vdots \\ \tau_9 \end{bmatrix} = X_1 \beta = 0$$

Es claro que la matriz X_1 tiene rango 1 i.e. el espacio engendrado por las columnas de X_1 es de dimensión 1 (o sea, la hipotesis H_1 tiene 1 g.d.l.).

El resultado de las pruebas de hipótesis están dadas en la siguiente tabla :

.../...

Fuente de variación	g.d.l.	S.C.	C.M.	F
TRATAMIENTO	8	157 448	19 681	1.30 NS(1)
H_1	1	57 207	57 207	3.78 NS
H_2	1	31 140	31 140	2.06 NS
H_3	3	25 693	8564 3	0.56 NS
H_4	3	43 408	14 469	0.95 NS
ERROR	36	544 690	15 130	

ANÁLISIS DEL MODELO 2

Se trata sucesivamente :

- las ecuaciones normales
- tabla de análisis de varianza
- estimación de los parámetros
- prueba de las hipótesis H_1, H_2, H_3, H_4

Las ecuaciones normales

u	β_1	β_2	β_3	β_4	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	Y	SUMAS
48	12	12	12	12	16	4	4	4	4	4	4	4	4	O	14 680
	12	O	O	O	4	1	1	1	1	1	1	1	1	87.13	4 383
		12	O	O	4	1	1	1	1	1	1	1	1	- 48.12	3 075
			12	O	4	1	1	1	1	1	1	1	1	16.80	4 752
				12	4	1	1	1	1	1	1	1	1	- 55.78	2 470
					16	O	O	O	O	O	O	O	O	- 5.01	5 858
						4	O	O	O	O	O	O	O	- 27.95	1 066
							4	O	O	O	O	O	O	- 31.20	1 265
								4	O	O	O	O	O	- 24.20	928
									4	O	O	O	O	10.05	877
										4	O	O	O	- 0.20	1 431
											4	O	O	13.55	1 241
												4	O	14.05	892
													4	66.05	1 122

t_{XX} (modelo 2)

t_{XX} (modelo 2)

.../...

(1) No significativo se denota NS

En la columna γ son las sumas de la covariable centrada asociadas a los diferentes modalidades de los factores.

TABLA DE ANÁLISIS DE VARIANZA

Fuente de variación	g.d.l.	S.C. covariable	SUMA DE PRODUCTOS	S.C. variable	MODELO 2		
					S.C.	C.M.	F
TRATAMIENTO	8	29 142	- 9 222	157 448	237 192	29 649	4.16**
BLOQUE	3	159 618	175 873	289 427	-	-	-
ERROR	36	121 408	189 278	544 690	249 601	7 131	
TRAT. + ERROR	44	150 550	180 056	702 138			

La prueba F del factor tratamiento en el modelo 2 es significativo

$$F = \frac{29649}{7131} = 4.16 \text{ **}$$

ESTIMACIÓN DE LOS PARAMETROS

La estimación de γ est $\hat{\gamma} = \frac{189278}{121408} = 1.559024$

Si se utilizan las mismas condiciones suplementarias entonces se estima :

μ por $Y_{...}$

τ_i por $Y_{i..} - Y_{...} - \hat{\gamma}(Z_{i..} - Z_{...})$

β_j por $Y_{.j.} - Y_{...} - \hat{\gamma}(Z_{.j.} - Z_{...})$

Testigo	τ_1	68.11
1 CN	τ_2	4.24
2 CN	τ_3	59.06
1 CS	τ_4	- 36.10
2 CS	τ_5	+102.25
1 CM	τ_6	+ 52.23
2 CM	τ_7	- 16.70
1 CK	τ_8	-104.73
2 CK	τ_9	-128.30

Prueba de las hipótesis H_1, H_2, H_3, H_4

Fuente de variación	g.d.l.	S.C. covariable	S. de productos	S.C.	g.d.l.	S.C.	Modelo 2 C.M.	F	
H_1	1	3 081	-13 276	52 276	1	103 465	103 465	14.51	**
H_2	1	2 204	8 285	31 140	1	104 473	10 473	1.46	NS
H_3	3	882	2 606	25 693	3	19 698	6 566	0.92	NS
H_4	3	22 975	- 6 837	43 408	3	332 607	110 869	15.54	+**
Error	36	121 408	189 278	544 690	35	249 601	7 131		
H_1 + error	37	124 489	176 002	601 897		353 066			
H_2 + error	37	123 612	197 563	575 830		260 074			
H_3 + error	39	122 290	191 884	570 383		269 299			
H_4 + error	39	144 383	182 441	588 098		357 567			

Debe notarse que :

1 - La suma de cuadrados residual en el modelo 2 es obtenida en base a los cálculos sobre Tratamientos + Error

$$702138 - \frac{(180056)^2}{150550} = 249601$$

2 - La suma de cuadrados ajustada correspondiente a la hipótesis H_3 por ejemplo es obtenida en base a los cálculos sobre H_3 + Error

$$570383 - \frac{(191884)^2}{122290} = 269299$$

después, restando la S.C. residual obtenemos :

$$269299 - 249601 = 19698$$

3 - La descomposición de la hipótesis global de la ausencia de efectos de los factores de los Tratamientos en H_1, H_2, H_3, H_4 son ortogonales al nivel del modelo (1)

$$157448 = 57207 + 31140 + 25693 + 43408$$

esta descomposición no es ortogonal al nivel del modelo 2 de covarianza

$$237192 \neq 103465 + 10473 + 19698 + 332607 = 466243$$

4 - Las hipótesis H_1 y H_4 son rechazadas
las hipótesis H_2 y H_3 no son rechazadas

TERCERO EJEMPLO : ESTUDIO DEL CRECIMIENTO DE LA VINA SEGUN LA PARTE QUE RECIBE EL INJERTO Y LA SOLUCION NUTRITIVA (1)

Nos referiremos aquí a los resultados de une experiencia sobre el crecimiento de la viña .

Se trata de estudiar el efecto de los factores cruzados "lugar de injerto" y "solución nutritiva" sobre el crecimiento de la viña midiendo el tamaño en diferentes fechas.

Factor "lugar injerto" : 4 niveles (4 lugares de injerto con vigores diferentes)
(i = 1,...,4)

Factor "solución nutritiva" : 3 niveles (3 soluciones diferentes)
(j = 1,...,3)

En cada una de las 12 células del plan de experiencia, 10 individuos han sido medidos en 29 fechas. hemos tomado aquí solamente datos referentes a 6 fechas : 15, 49, 70, 88, 105 y 137 días notados en la lista por ALT 15, ALT 49, ALT 70, ALT 88, ALT 105 y ALT 137.

lugar injerto	solución nutritiva	repetición	ALT 15	ALT 49	ALT 70	ALT 88	ALT 105	ALT 137
1	1	1	10.	68.	100.	120.	136.	158.
1	1	2	12.	69.	127.	159.	178.	202.
1	1	3	12.	80.	135.	170.	199.	228.
1	1	4	16.	101.	163.	200.	227.	257.
1	1	5	23.	109.	165.	201.	220.	237.
1	1	6	19.	106.	170.	208.	224.	238.
1	1	7	26.	108.	159.	184.	204.	228.
1	1	8	20.	93.	145.	164.	177.	196.
1	1	9	11.	60.	104.	129.	159.	196.
1	1	10	28.	114.	183.	224.	258.	293.

.../...

2	1	1	38.	121.	197.	246.	277.	314.
2	1	2	23.	99.	152.	188.	204.	226.
2	1	3	22.	94.	166.	214.	243.	277.
2	1	4	13.	66.	102.	121.	132.	148.
2	1	5	40.	130.	212.	266.	311.	351.
2	1	6	46.	133.	205.	247.	279.	306.
2	1	7	18.	79.	133.	151.	154.	161.
2	1	8	17.	90.	146.	173.	195.	224.
2	1	9	16.	83.	141.	187.	221.	261.
2	1	10	41.	141.	232.	280.	332.	376.
3	1	1	56.	117.	191.	216.	221.	246.
3	1	2	21.	101.	172.	214.	245.	285.
3	1	3	24.	100.	175.	225.	251.	286.
3	1	4	27.	112.	192.	238.	265.	304.
3	1	5	28.	109.	188.	234.	265.	301.
3	1	6	14.	77.	117.	144.	149.	159.
3	1	7	24.	92.	164.	204.	228.	247.
3	1	8	15.	95.	144.	175.	205.	236.
3	1	9	32.	118.	197.	237.	256.	281.
3	1	10	20.	108.	190.	248.	295.	339.
4	1	1	42.	134.	219.	259.	282.	306.
4	1	2	44.	140.	220.	270.	306.	341.
4	1	3	23.	113.	196.	259.	296.	322.
4	1	4	26.	110.	198.	250.	281.	315.
4	1	5	35.	138.	211.	267.	317.	361.
4	1	6	22.	112.	198.	250.	281.	308.
4	1	7	28.	110.	171.	210.	238.	268.
4	1	8	22.	104.	177.	223.	249.	276.
4	1	9	29.	112.	180.	222.	258.	291.
4	1	10	20.	97.	173.	222.	258.	291.
1	2	1	22.	72.	100.	110.	119.	131.
1	2	2	21.	75.	118.	144.	164.	187.
1	2	3	20.	77.	112.	132.	151.	174.
1	2	4	28.	78.	115.	143.	167.	184.
1	2	5	26.	87.	133.	163.	186.	214.
1	2	6	19.	83.	120.	133.	144.	159.
1	2	7	18.	66.	92.	105.	109.	112.
1	2	8	17.	72.	106.	124.	144.	161.
1	2	9	21.	65.	90.	109.	125.	142.
1	2	10	24.	91.	159.	209.	244.	275.

.../...

2	2	1	14.	58.	87.	99.	108.	117.
2	2	2	18.	71.	105.	127.	147.	171.
2	2	3	35.	97.	170.	214.	251.	283.
2	2	4	24.	83.	142.	174.	203.	231.
2	2	5	15.	55.	79.	92.	104.	117.
2	2	6	16.	74.	95.	106.	109.	110.
2	2	7	15.	59.	87.	100.	113.	129.
2	2	8	22.	75.	119.	156.	184.	212.
2	2	9	23.	79.	133.	174.	202.	212.
2	2	10	22.	75.	130.	167.	197.	230.
3	2	1	26.	80.	112.	132.	150.	170.
3	2	2	20.	62.	97.	116.	131.	150.
3	2	3	20.	69.	115.	138.	157.	182.
3	2	4	16.	65.	110.	141.	164.	187.
3	2	5	23.	87.	136.	160.	184.	206.
3	2	6	9.	62.	93.	108.	113.	113.
3	2	7	19.	94.	160.	208.	239.	273.
3	2	8	23.	89.	152.	198.	234.	268.
3	2	9	18.	76.	117.	151.	279.	209.
3	2	10	14.	54.	63.	69.	70.	70.
4	2	1	22.	84.	119.	137.	148.	159.
4	2	2	23.	75.	113.	133.	144.	157.
4	2	3	19.	71.	121.	158.	189.	220.
4	2	4	33.	108.	174.	220.	245.	274.
4	2	5	41.	118.	186.	239.	277.	317.
4	2	6	32.	85.	103.	112.	113.	113.
4	2	7	45.	112.	180.	219.	252.	263.
4	2	8	31.	100.	159.	197.	231.	258.
4	2	9	24.	109.	197.	245.	288.	320.
4	2	10	30.	112.	209.	268.	312.	342.
1	3	1	23.	91.	133.	154.	166.	179.
1	3	2	16.	57.	87.	99.	107.	115.
1	3	3	13.	53.	72.	82.	86.	86.
1	3	4	28.	79.	110.	125.	130.	132.
1	3	5	19.	70.	120.	151.	174.	198.
1	3	6	17.	64.	102.	126.	146.	150.
1	3	7	19.	69.	87.	92.	92.	92.
1	3	8	25.	79.	123.	157.	185.	216.
1	3	9	31.	99.	149.	173.	190.	203.
1	3	10	23.	100.	157.	194.	224.	248.

.../...

2	3	1	23.	70.	101.	113.	119.	126.	
2	3	2	22.	78.	128.	164.	192.	218.	
2	3	3	27.	97.	153.	198.	223.	249.	
2	3	4	30.	87.	121.	140.	154.	164.	
2	3	5	20.	66.	104.	125.	136.	149.	
2	3	6	26.	83.	142.	186.	219.	244.	
2	3	7	26.	98.	153.	182.	201.	219.	
2	3	8	16.	74.	124.	165.	202.	238.	
2	3	9	33.	102.	166.	216.	248.	283.	
2	3	10	22.	57.	118.	154.	181.	211.	
3	3	1	25.	77.	126.	161.	182.	209.	
3	3	2	15.	77.	127.	156.	176.	202.	
3	3	3	26.	96.	146.	172.	187.	210.	
3	3	4	18.	83.	132.	166.	191.	219.	
3	3	5	27.	110.	189.	233.	262.	295.	
3	3	6	14.	58.	96.	116.	130.	150.	
3	3	7	27.	96.	140.	167.	192.	219.	
3	3	8	16.	69.	117.	150.	173.	197.	
3	3	9	19.	67.	105.	134.	155.	175.	
3	3	10	26.	100.	142.	179.	211	241.	
4	3	1	36.	106.	153.	175.	188.	205.	
4	3	2	21.	86.	125.	148.	163.	177.	
4	3	3	26.	96.	159.	187.	211.	234.	
4	3	4	24.	92.	160.	199.	232.	249.	
4	3	5	22.	84.	123.	147.	159.	165.	
4	3	6	31.	103.	176.	225.	251.	272.	
4	3	7	26.	97.	143.	177.	186.	198.	
4	3	8	19.	71.	134.	163.	191.	213.	
4	3	9	49.	110.	184.	244.	283.	319.	
4	3	10	41.	142.	213.	253.	276.	289.	

Así hemos analizado los datos a través del modelo de análisis de varianza multivariable.

$$Y_{ijk}^{(t)} = \mu^{(t)} + \alpha_i^{(t)} + \beta_i^{(t)} + \theta_{ij}^{(t)} + \varepsilon_{ijk}^{(t)} \qquad (1)$$

con :

$Y_{ijk}^{(t)}$: altura al instante $t(t=1,\ldots,6)$ de la $k^{ésima}$ repetición de la celda (i,j) del plan de experiencia $(k=1,\ldots,10)$

$\mu^{(t)}$: media general al instante t

$\alpha_i^{(t)}$: efecto "medio" al instante t del nivel i del factor "lugar-injerto" $(i=1,\ldots,4)$

$\beta_j^{(t)}$: efecto "medio" al instante t del nivel j del factor "solución nutritiva" $(j=1,\ldots,3)$

$\theta_{ij}^{(t)}$: interacción al instante t entre lugar de injerto i y solución nutritiva j

.../...

Las variables $\varepsilon_{ijk}^{(t)}$ se consideran :

- independientes de un individuo al otro
- dependientes a nivel del individuo en el tiempo, de matriz de covarianza

$$\text{cov} \begin{bmatrix} \varepsilon_{ijk}^{(1)} \\ \cdot \\ \cdot \\ \varepsilon_{ijk}^{(6)} \end{bmatrix} = \underset{(6,6)}{\Sigma}$$

de estructura independiente del individuo (i,j,k)

- normalmente distribuidas.

Hemos efectuado un análisis de varianza multivariado sobre el conjunto de 6 medidas de altura para preparar las pruebas multivariadas y el análisis discriminante que seguirán.

Prueba de hipótesis : ausencia de interacción

Podemos formalizar el problema de la forma siguiente :

$$H_o : \forall\ (i,j),\ t \quad ; \ \theta_{ij}^{(t)} \cdot = 0$$

contra $$\bar{H}_o : \exists\ (i,j),\ t \quad ; \ \theta_{ij}^{(t)} \neq 0$$

La estadística utilizada para la prueba es la del valor propio más grande de la matriz INTER. En este caso, el valor propio más grande no es significativo (χ^2 = 27.007 con 36 g.d.l.). Se conluye por lo tanto que no hay efecto de interacción θ.

Los resultados del análisis son los siguientes :

No de g.d.l. INTER = 6

No de g.d.l. INTRA = 108

MATRIZ INTER CON RESPECTO A LAS VARIABLES REDUCIDAS

1	1.451					
2	.850	.600				
3	.807	.549	.547			
4	.698	.496	.513	.505		
5	.562	.438	.454	.465	.445	
6	.499	.433	.452	.473	.469	.507

.../...

No de valores proprios = 6

VAL. PROPRIO	G.D.L.	CHI^2	SIGNIFICACION	HIPOTESIS
2.245	36	27.007	NS	Valor propio N°1
1.366	25	14.370	NS	Valor propio N°2
.548	16	6.509	NS	Valor propio N°3
.493	9	3.285	NS	Valor propio N°4
.062	4	.382	NS	Valor propio N°5
.002	1	.015	NS	Valor propio N°6

VAL. PROPRIO * G.D.L. INTER	G.D.L.	SIGNIFICACION
.015	1	.903
.370	3	.946
2.956	5	.706
3.288	7	.857
8.193	9	.514
13.472	11	.263

Podemos efectuar separadamente un análisis discriminante sobre el factor "lugar injerto" y sobre el factor "solución nutritiva".

Análisis factorial discriminante sobre el factor "Lugar Injerto"

En los resultados del análisis encontramos un efecto "lugar-injerto" muy marcado. La estadística de Fisher correspondiente a la primera variable conónica es igual a F = $_1$ = 17.16 (contra F=12.961 para ALT 70 a ver sobre la matriz INTER con respecto a las variables reducidas).

En la búsqueda del espacio de representación se obtiene por lo tanto un espacio de dos dimensiones para cuatro puntos (2 valores propios significativos).

Tratemos de interpretar los dos primeros vectores canónicos (expresados en la base de variables incidados reducidas INTRA). Para hacerlo, veamos las correlaciones en el sentido INTER entre las variables iniciales y las dos variables canónicas (1).

La 1er variable esta muy correlacionada positivamente con todas las variables "altura" y proporciona una escala de "vigor".

.../...

(1) Las variables canónicas pueden ser llamado componentes principales discriminante.

La 2da. variable canónica tiene correlaciones positivas pequeñas con las variables iniciales (la correlación augmenta con el tiempo hasta el valor 0.358). No encontramos una interpretación clara (note que $\lambda_2 = 6.521$ es inferior al más pequeño F = 8.09 del análisis de varianza para ALT 15).

No de g.d.l. INTER = 3
No de g.d.l. INTRA = 108

MATRIZ INTER CON VARIABLES REDUCIDAS

	1	2	3	4	5	6
1	8.095					
2	9.250	11.377				
3	9.920	12.025	12.961			
4	9.796	11.753	12.784	12.664		
5	9.191	10.926	11.922	11.832	11.069	
6	8.325	9.854	10.876	10.846	10.160	9.385

- BUSQUEDA DE LA DIMENSION DEL ESPACIO DE REPRESENTACION

No de valores propios= 3

VAL. PROPIO	D.L.	CHI2	SIGNIFICACION	HIPOTESIS
17.159	18	67.177	**	Valor propio No 1
6.521	10	25.861	**	Valor propio No 2
2.901	4	8.215	NS	Valor propio No 3

VAL. PROPIO±	G.D.L. INTER	G.D.L.	SIGNIFICACION
8.703		4	0.0690
19.562		6	0.0033
51.477		8	0.0000

VECTORES PROPIOSEN LA BASE DE VECTORES PROPIOS DE W

	1	2	3
1	.8286	.3242	.0288
2	- .1236	.3071	- .0983
3	- .1072	.2214	.9576
4	- .3176	-.1420	.1125
5	- .3599	.1446	- .1662
6	- .2373	.8429	- .1794

.../...

VECTORES PROPIOS EN LA BASE INICIAL REDUCIDA

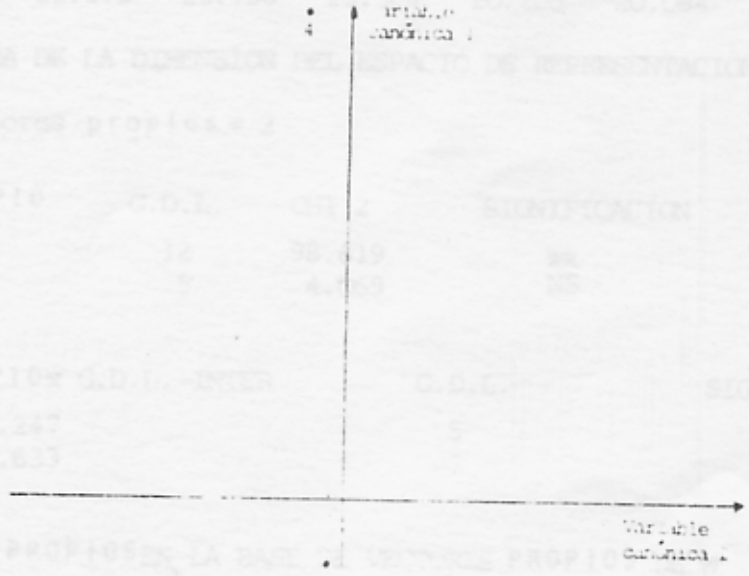
	1	2	3
1	.2566	-.1703	1.465
2	-.1323	-.5705	- 1.425
3	-.5713	-1.037	- 1.690
4	.7814	9.721	- 1.078
5	5.665	-16.35	1.494
6	-5.203	8.713	- 1.494

CORRELACIONES ENTRE VARIABLES CANONICAS E INICIALES EN EL SENTIDO INTRA

	1	2	3
1	.676	.038	.272
2	.805	.148	- .206
3	.845	.328	- .075
4	.825	.391	.017
5	.769	.373	.083
6	.689	.429	.120

CORRELACIONES ENTRE VARIABLES CANONICAS E INICIALES EN EL SENTIDO INTER

	1	2	3
1	.984	.079	.163
2	.988	.112	-.104
3	.972	.233	-.036
4	.960	.280	.008
5	.957	.286	.043
6	.931	.358	.067



.../...

Análisis factorial discriminante sobre el factor "solución nutritiva"

En el análisis hay un efecto "solucion nutritiva" muy significativo. La estadística de FISHER correspondiente a la 1ª variable canónica es igual a : $F = \lambda_1 = 78.31$ (contra F = 28.2221 para ALT 70).

Se obtiene un espacio de reoresentación de 3 puntos en un espacio de una dimension (1 valor proprio significativo).

Aquí la 1ª variable canónica está muy correlacionada (negativamente) con las variables iniciales y da una "escala" de la acción de las soluciones sobre la altura. Se nota que la solución 1 es muy diferente de las soluciones 2 y 3.

No. de g.d.l. INTER = 2
No. de g.d.l. INTRA = 108

MATRIZ INTER CON LAS VARIABLES REDUCIDAS

	1	2	3	4	5	6
1	.891					
2	3.941	22.474				
3	4.374	25.181	28.222			
4	4.118	23.705	26.568	25.011		
5	3.648	21.366	23.960	22.556	20.362	
6	3.563	21.175	23.756	23.364	20.205	20.064

- BUSQUEDA DE LA DIMENSION DEL ESPACIO DE REPRESENTACION

No de valores propios = 2

VAL PROPIO	G.D.L.	CHI 2	SIGNIFICACION	HIPOTESIS
78.316	12	98.619	**	Valor propio No 1
2.123	5	4.069	NS	Valor propio No 2

VAL PROPIO*	G.D.L. INTER	G.D.L.	SIGNIFICACION
4.247		5	0.5144
156.633		7	0.0000

VECTORES PROPIOS EN LA BASE DE VECTORES PROPIOS DE W

	1	2
1	- .5184	- .1205
2	- .2961	.3849
3	.5369	.0019
4	.2145	.5316
5	- .2663	.7008
6	- .4882	- .2521

VECTORES PROPIOS EN LA BASE INICIAL REDUCIDA

	1	2
1	1.094	- 5.288
2	- .4223	.2668
3	- 1.738	4.063
4	- 4.487	- 9.870
5	10.48	2.676
6	- 5.218	3.315

CORRELACIONES ENTRE VARIABLES CANONICAS E INICIALES EN EL SENTIDO INTER

	1	2
1	-.091	-. 338
2	-. 535	-. 176
3	-. 600	-. 133
4	-. 565	-. 125
5	-. 510	-. 016
6	-. 506	. 064

CORRELACIONES ENTRE VARIABLES CANONICAS E INICIALES EN EL SENTIFO INTRA

	1	2
1	-. 854	-. 521
2	-. 999	-. 054
3	-. 999	-. 037
4	-. 999	-. 037
5	-1.000	-. 005
6	-1.000	. 021

.../...

CUARTO EJEMPLO : ANALISIS DEL PORCENTAGE DE PROTEINAS EN 25 VARIEDADES DE CEBADA CULTIVADOS EN PARCELAS AL AZAR SOBRE 3 BLOQUES DE UN DISEÑO EN BLOQUES COMPLETOS (1)

En este ejemplo se trata principalmente de explicar el porcentaje de proteínas en función de los factores del plan de experiencia(factores:"variedad"y"bloque")y de la covariable "rendimiento". Aquí la covariable es interna al proceso experimental, es decir, que no está determinada antes de la experiencia. Así el problema es de estudiar el porcentaje de proteínas condicionalmente al rendimiento. Sin dar al condicionamiento un sentido abusivo, diremos aquí que la covariable sirve únicamente para enmendar el plan de experiencia.

En este caso, la ecuación es

$$Y_{ij} = \mu + v_i + b_j + ax_{ij} + \varepsilon_{ij}$$

con :

Y_{ij} : % proteínas
v_i : efecto variedad i (i=1,...,25)
b_j : efecto bloque j (j=1,2,3)
x_{ij} : covariable rendimiento
a : coeficiente de regresión de la covariable rendimiento.

Otra manera de atacar el problema es utilizar el Análisis de Varianza Multivariable.

En esta caso las ecuaciones son:

$$Y_{ij} = \mu_1 + v_{1j} + b_{1j} + \varepsilon_{1ij}$$

$$X_{ij} = \mu_2 + v_{2i} + b_{2j} + \varepsilon_{2ij}$$

donde v_{1i} , b_{1j} , y v_{2i} , b_{2j} son factores variedad y bloque en las dos ecuaciones arriba citadas.

En la exposición se encontrará:

lista de datos

análisis de varianza multivariable

análisis de covarianza

Variedad bloque		1	2	3	4	5	6	7	8	9	10	11	12
1	% PROTEINAS	15.80	17.80	15.00	16.15	15.35	16.50	14.85	14.65	15.05	15.70	14.65	14.25
	RENDIMIENTO	32.70	30.10	39.50	35.20	43.20	36.00	44.20	48.00	40.00	40.00	40.60	50.00
2	% PROTEINAS	15.75	18.40	14.50	15.10	14.60	15.65	14.80	15.40	15.20	15.85	14.05	14.55
	RENDIMIENTO	31.40	31.40	39.00	37.40	42.40	39.80	42.80	41.00	44.00	39.40	47.80	47.20
3	% PROTEINAS	14.85	17.90	14.75	15.80	14.90	15.50	15.45	15.45	15.35	15.60	14.20	14.75
	RENDIMIENTO	37.30	32.20	38.20	38.80	38.40	43.20	40.00	43.80	44.60	42.50	45.40	38.00

Variedad / Bloque		13	14	15	16	17	18	19	20	21	22	23	24	25
1	% PROTEINAS	14.55	14.85	15.50	16.60	14.95	14.85	15.20	14.30	14.50	14.00	14.85	14.70	14.60
	RENDIMIENTO	41.00	42.00	33.00	34.80	38.80	40.60	35.80	40.80	49.80	42.40	42.20	46.80	42.80
2	% PROTEINAS	15.15	14.80	15.70	15.40	15.60	14.35	16.15	13.25	14.45	13.90	14.85	15.30	14.00
	RENDIMIENTO	40.40	42.20	34.40	37.60	33.20	40.20	31.20	40.40	46.40	50.60	47.20	48.60	46.00
3	% PROTEINAS	15.25	14.70	15.80	15.70	15.65	14.90	15.75	14.65	14.15	13.66	15.30	15.60	14.70
	RENDIMIENTO	40.20	45.40	35.60	41.40	34.00	37.20	36.00	35.20	48.20	46.60	42.40	44.80	41.80

Análisis de varianza

CARACTER 1 (RENDIMIENTO)

Fuente de variación	S.C.	G.D.L.	C.M.	F
VARIEDAD	$.144632 \times 10^4$	24	$.6026 \times 10^2$	7.485 **
BLOQUE	$.3391 \times 10^1$	2	$.169 \times 10^1$	.211
ERROR	$.38647 \times 10^3$	48	$.805 \times 10^1$	
TOTAL	$.183618 \times 10^4$	74		

CARACTER 2 (PROTEINA)

Fuente de variación	S.C.	G.D.L.	C.M.	F
VARIEDAD	$.4939 \times 10^2$	24	$.2057 \times 10^1$	13.465 **
BLOQUE	.2654	2	.1327	.868
ERROR	$.7336 \times 10^1$	48	.1528	
TOTAL	$.569 \times 10^2$	74		

MATRIZ DE CORRELACIÓN INTER DEL FACTOR No1

	Rendimiento	% Proteinas
Rendimiento	1	
% Proteinas	- .693	1

MATRIZ DE CORRELACIÓN INTER DEL FACTOR No2

	Rendimiento	% Proteinas
Rendimiento	1	
% Proteinas	- .929	1

MATRIZ DE CORRELACIÓN INTRA

	Rendimiento	% Proteinas
Rendimiento	1	
% Proteinas	- .554	1

MATRIZ DE CORRELACIÓN INTER GLOBAL

	Rendimiento	% Proteinas
Rendimiento	1	
% Proteinas	- .693	1

VARIABLES ANALIZADAS	Rendimiento	% Proteinas
MEDIA GENERAL	$.4058 \times 10^2$	$.1515 \times 10^2$

MEDIAS DEL FACTOR VARIEDAD

2	31.23	18.03
6	39.67	15.88
16	37.93	15.80
10	40.63	15.72
19	34.33	15.70
4	37.13	15.68
15	34.33	15.67
1	33.80	15.47
17	35.33	15.40
9	42.87	15.20
24	46.73	15.20
8	44.27	15.17
7	42.33	15.03
23	43.93	15.00
13	40.53	14.98
5	41.33	14.95
14	43.20	14.78
3	38.90	14.75
18	39.33	14.70
12	45.07	14.52
25	43.53	14.43
21	48.13	14.37
11	44.60	14.30
20	38.80	14.07
22	46.53	13.85

MEDIAS DEL FACTOR BLOQUE

3	40.45	15.21
1	40.41	15.17
2	40.88	15.07

.../...

Aquí los niveles de los factores variedad y bloque son clasificados según valores decrecientes del factor porcentaje de proteínas. Se puede ver que entre más grandes son los rendimientos, más pequeños son los porcentajes de proteínas. Sin embargo, esta afirmación es falsa para algunas variedades. Por consiguiente, se necesita clasificar las variedades tomando en cuenta conjuntamente las dos variables.

ANÁLISIS DE VARIANZA MULTIVARIABLE SOBRE EL FACTOR VARIEDAD

No de g.d.l. INTER = 24
No de g.d.l. INTRA = 48

MATRIZ INTER CON VARIABLES REDUCIDAS

	Rendimiento	% Proteinas
Rendimiento	7.4848	
% Proteinas	- 6.9538	13.4646

- BUSQUEDA DE LA DIMENSION DEL ESPACIO DE REPRESENTACIÓN

No de VALORES PROPIOS = 2

VAL.. PROPIO	G.D.L.	CHI.2	SIGNIFICACION	HIPÓTESIS
13.49	48	197.24	**	Valor propio 1
5.52	23	77.49	**	Valor propio 2

- CORRELACIONES ENTRE VARIABLES CANÓNICAS E INICIALES EN EL SENTIDO INTER

	Variable canónica 1	Variable canónica 2
Rendimiento	.666	.745
% Proteínas	- .999	- .035

En este análisis multivariable, se tiene un efecto variedad más grande que en el análisis univariado. La estadística de FISHER correspondiente a la primera variable canónica es igual a 13.49 contra 13.46 en el análisis de la variable porcentaje de proteínas en el caso univariado.

.../...

Así se obtiene un espacio de representación de los 25 puntos medios de dimensión 2 (dos valores propios son significativos).

La primera variable canónica tiene correlación positivamente con el rendimiento y negativamente con el porcentaje de proteí - nas. La segunda variable canónica tiene correlación únicamente con el rendimiento.

No es importante aquí dar una significación a las variables canónicas. Más importante es ver cómo están representadas las variables en el espacio canónico. (gráfico 1).

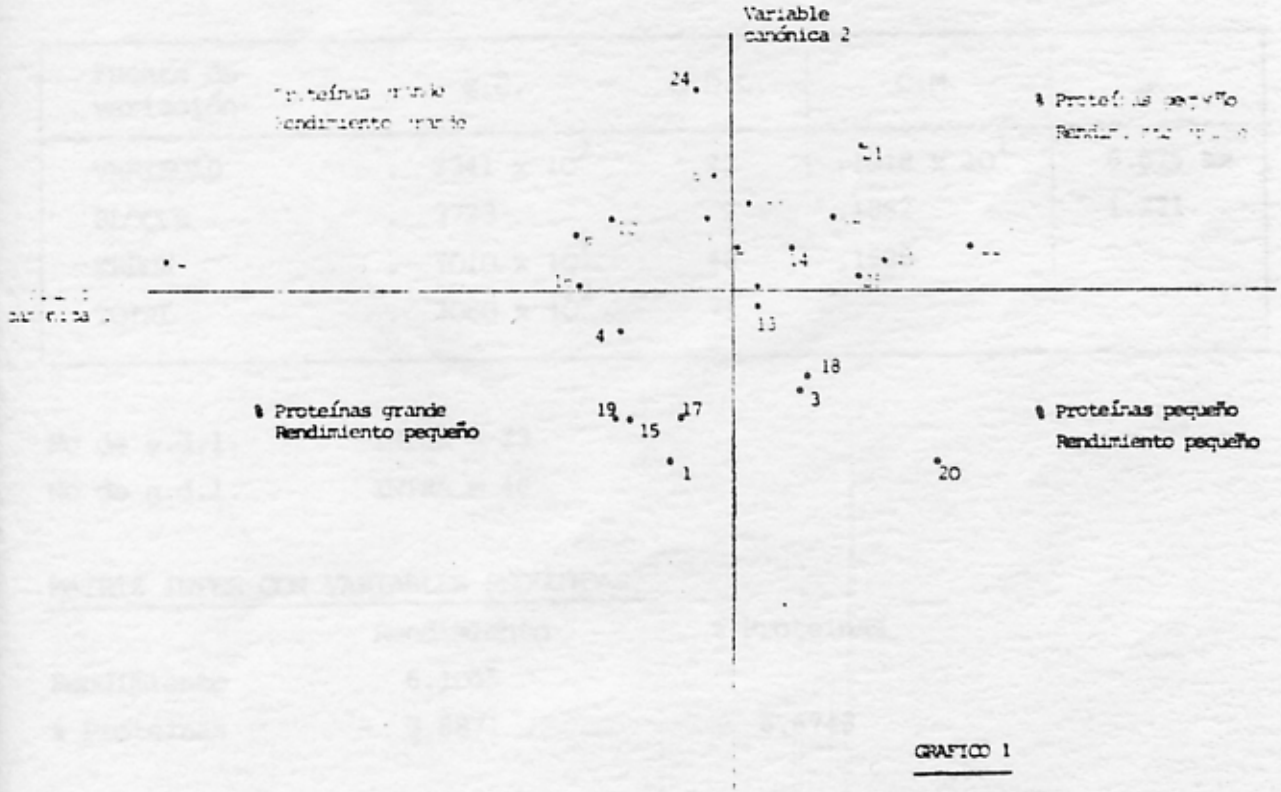


GRAFICO 1

.../...

En el gráfico anterior parece que la variedad 2 es particular. Así hemos hecho otro análisis sin la variedad 2. Los resultados son :

CARACTER 1 (RENDIMIENTO)

Fuente de variación	S.C.	G.D.L.	C.M.	F
VARIEDAD	$.117332 \times 10^4$	23	$.5101 \times 10^2$	6.106 **
BLOQUE	$. 3391 \times 10^1$	2	$.1696 \times 10^1$	0.203
ERROR	$. 38435 \times 10^3$	46	$.8355 \times 10^1$	
TOTAL	$.156106 \times 10^4$	71		

CARACTER 2 (PROTEINA)

Fuente de variación	S.C.	G.D.L.	C.M.	F
VARIEDAD	$. 2341 \times 10^2$	23	$.1018 \times 10^1$	6.675 **
BLOQUE	. 3723	2	.1862	1.221
ERROR	$. 7010 \times 10^1$	46	.1525	
TOTAL	$. 3080 \times 10^2$	71		

No de g.d.l. INTER = 23

No de g.d.l. INTRA = 46

MATRIZ INTER CON VARIABLES REDUCIDAS

	Rendimiento	% Proteinas
Rendimiento	6.1055	
% Proteinas	- 3.8871	6.6746

- BUSQUEDA DE LA DIMENSIÓN DEL ESPACIO DE REPRESENTACIÓN

No de VALORES PROPRIOS = 2

VAL. PROPRIO	G.D.L.	CHI 2	SIGNIFICACION	HIPOTESIS
6.702	46	156.96	**	Valor proprio 1
5.581	22	74.62	**	Valor proprio 2

.../...

CORRELACIONES ENTRE VARIABLES CANONICAS E INICIALES EN EL SENTIDO INTER

	Variable canónica 1	Variable canónica 2
Rendimiento	- 0.716	0.697
% Proteinas	0.989	0.143

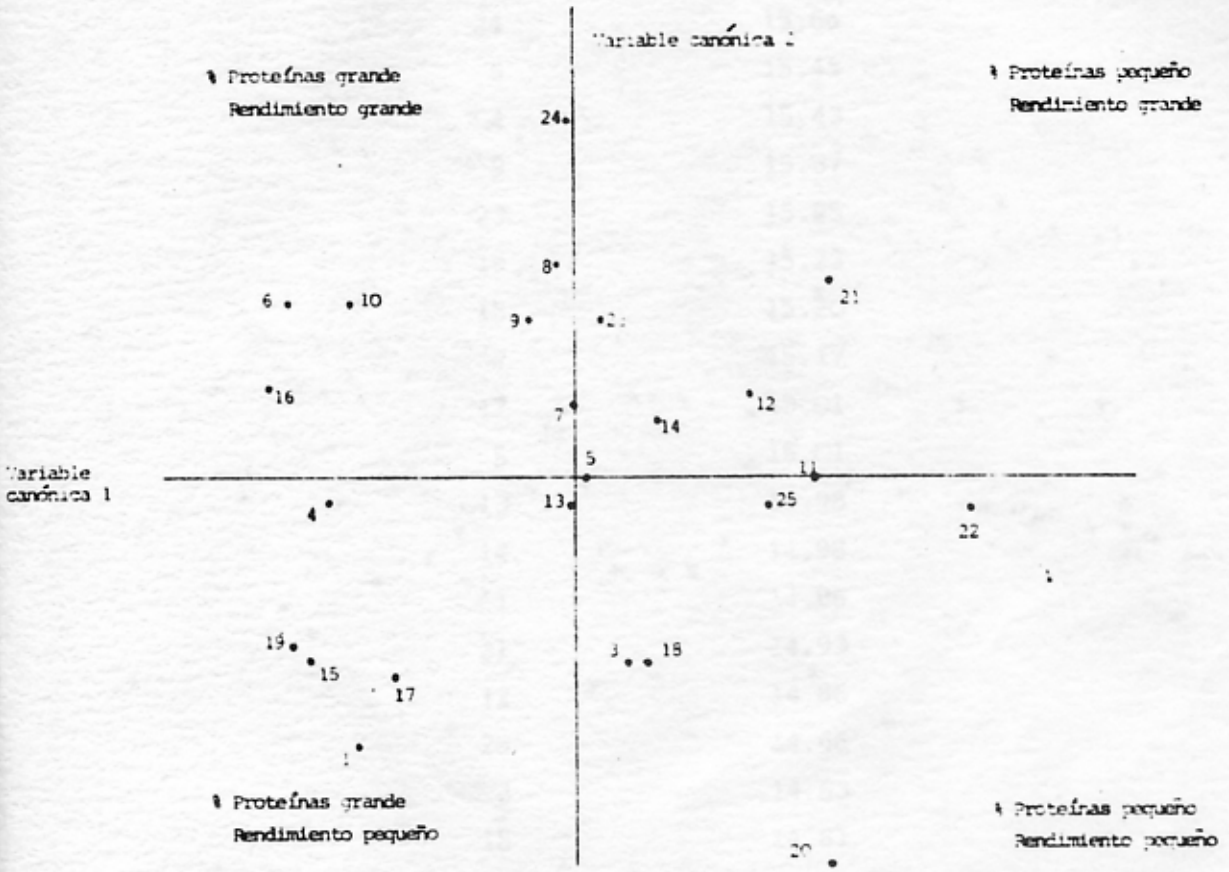


GRAFICO 2

Análisis de covarianza según el modelo

$$Y_{ij} = \mu + v_i + b_j + ax_{ij} + e_{ij}$$

CARACTER (PROTEINA)

Fuente de variación	S.C.	G.D.L.	C.M.	F
VARIEDAD	$.2654 \times 10^2$	24	.11509	10.094 **
BLOQUE	.1510	2	. 0755	0.689
COVARIABLE	.2186	1	. 2186	19.96 **
ERROR	.5149 x 10	47	. 1096	
TOTAL	$.5699 \times 10^2$	74		

B I B L I O G R A F I A

(1) BADIA J., MASSON J.P. (1977) - Analyse de variance - Publication du Département de Biométrie.

(2) COCHRAN W.G., COX G.M. (1957) - Experimental design - John Wiley & Sons.

(3) CROQUETTE A., PASTOR J., SCHEKTMAN Y. (1980) - Etude descriptive du "dynamisme" d'une population d'entreprises industrielles, à partir d'un indice base sur les consommations mensuelles d'électricité. Application de l'analyse en composantes principales à des données temporelles.
Université Paul Sabatier - Toulouse.

(4) DENIS J.B. (1976) - Tratamiento de datos - I.N.I.A. MADRID.

(5) RAO C.R. (1967) - Linear statistical inference and its application - John Wiley & Sons.

(6) SANTANA E. (1979) - Logement des travailleurs émigrés à Toulouse. Une étude sur la localisation -ségrégation sur l'espace urbain.
Thèse de 3ième cycle Université du Mirail-Toulouse.

(7) SCHEKTMAN Y. (1978) - Analyse en composantes M-semblables - Université Paul Sabatier - Toulouse.

(8) SCHEKTMAN Y. (1978) - Contribution à la mesure en facteurs dans les sciences expérimentales et à la mise en oeuvre automatique des calculs statistiques.
Thèse de docteur d'Etat-Sciences - Université Paul Sabatier - Toulouse.

(9) SEARLE S.R. (1971) - Linear models - New York - John Wiley & Sons.

MEDIAS DEL FACTOR VARIEDAD	
2	17.33
6	15.82
10	15.72
16	15.70
24	15.66
8	15.45
4	15.43
9	15.37
23	15.25
19	15.23
15	15.20
7	15.17
17	15.01
5	15.01
13	14.98
14	14.98
1	14.96
21	14.93
12	14.86
25	14.66
3	14.63
18	14.61
11	14.60
22	14.30
20	13.94

VALOR DEL COEFICIENTE DE REGRESION a = - 0.07522

Si se comparan las clasificaciones de las variedades en los tres análisis, vemos que las clasificaciones obtenidas en el análisis de varianza multivariable (en proyección sobre el primer eje canónico) y en el análisis de covarianza son idénticos pero diferentes de la calsificación obtenida en el análisis de varianza univariable.

El análisis de varianza multivariable es el más rico porque la clasificación obtenida es en el plano de las variables canónicas y diferentes partes del plano pueden ser identificadas.