ANÁLISIS DISCRIMINANTE

DEFINICIÓN:

- **Cómo técnica de análisis de dependencia:**

Pone en marcha un modelo de causalidad en el que la variable endógena es una variable NO MÉTRICA y las independientes métricas.

- **Cómo técnica de análisis de clasificación:**

Ayuda a comprender las diferencias entre grupos. Explica, en función de características métricas observadas, porqué los objetos/sujetos se encuentran asociados a distintos niveles de un factor.

DIFERENCIAS CON.....:

- **El análisis de regresión:** En la regresión, la endógena es métrica

- **El análisis ANOVA:** En el ANOVA, la endógena es métrica y las exógenas NO MÉTRICAS (al contrario que en el discriminante)

- **El LOGIT - PROBIT:** Idéntica al discriminante en el objetivo pero apoyada en técnicas de estimación paramétrica idénticas a la regresión y no en análisis de descomposición de la varianza:

    (1) **DV:** Más adecuada para factores sólo binarios
    (2) **DV:** Más compleja de cálculo - interpretación

    (3) **V:** Se ve menos afectada por incumplimientos de supuestos teóricos necesarios a priori (normalidad, por ejemplo)
    (4) **V:** Permite incorporar explicativas no métricas en forma de ficticias
    (5) Los resultados admiten explotación en términos de probabilidad

ANÁLISIS DISCRIMINANTE

QUÉ SIGNIFICA, EN ESTE CONTEXTO.....:

- **ADP y ADD:** Análisis Discriminante **Descriptivo** o **Predictivo**
- **MDA:** Análisis discriminante **Múltiple** (no binario)

A. DISCRIMINANTE DESCRIPTIVO (Un ejemplo):

(Objetivo) Se desea caracterizar el perfil de los compradores de un determinado producto en un determinado establecimiento.

(Diseño) Para ello, se diseña una muestra con 100 compradores y 100 no compradores y se toman datos de renta, edad y cercanía al establecimiento de venta.

(Resultado) El análisis discriminante establecerá la importancia relativa de cada uno de estos atributos en la decisión de compra permitiendo orientar mejor la política promocional o de distribución del producto.

B. DISCRIMINANTE DESCRIPTIVO (otro ejemplo):

(Objetivo) Se desea valorar de qué depende la fidelidad de un clientes a un determinado proveedor comercial.

(Diseño) Para ello, se encuesta a 15 importantes clientes sobre la posibilidad de cambiar de proveedor y sobre la percepción que estos tienen de su Competitividad y Nivel de Servicio.

(Resultado) El análisis permitirá aproximar la importancia relativa de la competitividad y el nivel de servicio a la hora de conseguir fidelidad en un cliente.

C. DISCRIMINANTE PREDICTIVO (un ejemplo):

(Objetivo) Se desea prever el riesgo de morosidad relativa a los préstamos personales en una entidad bancaria.

(Diseño) Se explota el fichero histórico de clientes morosos - no morosos y se observan variables cuantitativas potencialmente explicativas: renta total, edad, créditos adicionales, años de estabilidad laboral, ...

ANÁLISIS DISCRIMINANTE

(Resultado) Aplicando el modelo estimado con el fichero histórico, el análisis permitirá anticipar el riesgo de morosidad de nuevos clientes.

ETAPAS PARA DE UN ANÁLISIS DISCRIMINANTE

A.- SELECCIÓN DE VARIABLES DEPENDIENTE E INDEPENDIENTES

B.- SELECCIÓN DEL TAMAÑO MUESTRAL

C.- DIVISIÓN DE LA MUESTRA

D.- CHEQUEO DE LAS HIPÓTESIS DE PARTIDA

E.- ESTIMACIÓN DEL MODELO

F- VALIDACIÓN DE LAS FUNCIONES DISCRIMINANTES

G.- CONTRIBUCIÓN DE LAS VARIABLES A LA CAPACIDAD DISCRIMINANTE DE LAS FUNCIONES

H.- VALORACIÓN DE LA CAPACIDAD PREDICTIVA

I.- UTILIZACIÓN FUNCIONES

ANÁLISIS DISCRIMINANTE

A.- SELECCIÓN DE VARIABLES DEPENDIENTE E INDEPENDIENTES

- La **variable dependiente** no tiene que ser, necesariamente, categórica en origen

- Los grupos deben ser **mutuamente excluyentes**

- La decisión sobre **el número de categorías**

    (1) debe ajustarse al poder discriminante de los predictores
    (2) puede observarse en etapas sucesivas (*inicial con todas, y en el límite, optando sólo por el enfoque de extremos polares)*

- Las **variables explicativas:**

    (1) no deben ser excesivas
    (2) deben atender siempre al objetivo conceptual
    (3) pueden someterse a un test univariante de diferencia de medias o un test ANOVA

B.- SELECCIÓN DEL TAMAÑO MUESTRAL

- Elevada sensibilidad al tamaño muestral Vs. Nº de predictoras. *(Receta: mínimo 5 observaciones por variable..... recomendado 20 observaciones por variable).*

- También debe vigilarse el tamaño de los grupos:

    (1) el equilibrio no es necesario pero es recomendable
    (2) el más pequeño de los grupos no puede serlo mucho (*Receta: como mínimo, el tamaño del grupo más pequeño debe ser mayor al número de variables*).

ANÁLISIS DISCRIMINANTE

C.- DIVISIÓN DE LA MUESTRA

- Utilidad del "enfoque de validación cruzada (*muestra de análisis* + *muestra ampliada*)"

    (1) garantizado un tamaño muestral total suficiente
    (2) aplicando muestreo estratificado proporcional en ambas muestras

D.- CHEQUEO DE LAS HIPÓTESIS DE PARTIDA

- Ausencia de normalidad multivariante ⇒ problemas en la estimación ⇒ LOGIT recomendado

- Matrices de varianzas y covarianzas distintas ⇒ problemas en la clasificación ⇒ uso de técnicas de clasificación cuadráticas

- Multicolinealidad ⇒ problemas en la interpretación de parámetros ⇒ estimación secuencial

E- ESTIMACIÓN DEL MODELO

SELECCIÓN DEL MÉTODO (I)

- Método simultáneo o por etapas:

    (1) **estimación en una sola etapa** (número reducido de variables, interés por el conjunto)

    (2) estimación polietápica: selección de menos a más, analizando las interacciones de las variables discriminantes (*amplio número de variables, dudas sobre el modelo teórico*)

- Método cálculo : Método de Fisher, D de Mahalanobis, ….

ANÁLISIS DISCRIMINANTE

SELECCIÓN DEL MÉTODO (II)

(Noción básica sobre el método de Fisher)

Elemento	Grupo	**X1**	**X2**	**X3**		**Y**=f(x1,x2,x3)
1	A	**25**	**25**	**23**	→	10
2	A	**15**	**14**	**26**		12
3	A	**14**	**13**	**21**	→	11
4	B	**25**	**18**	**41**		5
5	B	**65**	**14**	**18**	→	4
6	B	**15**	**18**	**48**		4

- Variable Y (Función Discriminante): combinación lineal de las variables originales "X" que:

  (1) Presente la mínima variación INTRA grupal
  (2) Presente la máxima variación ENTRE grupal

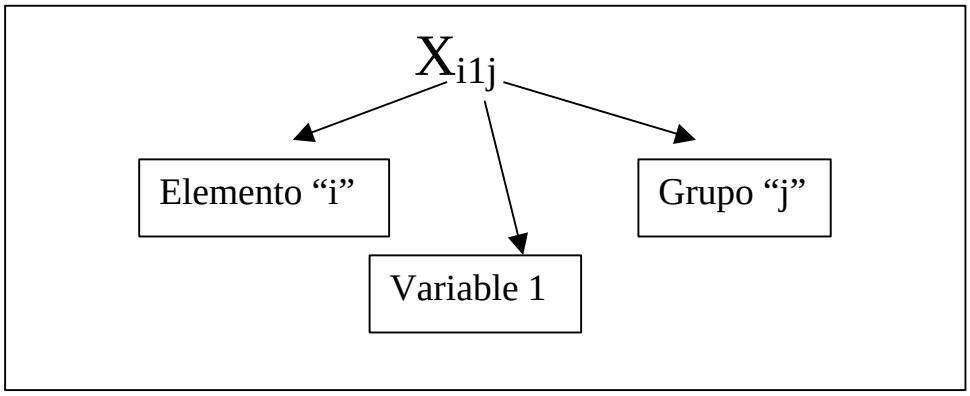
- La función discriminante no será única: si se parte de una clasificación en "g" grupos, se obtendrán varios conjuntos de parámetros, es decir, varias funciones discriminantes (Menor de “g-1” o “p”)

ANÁLISIS DISCRIMINANTE

SELECCIÓN DEL MÉTODO (II) *(Continuación)*

(Obtención de la/s funciones discriminantes)
(Planteamiento para "g" grupos y "p" variables)

- **Partimos de:**
  - **"g"** grupos que representamos con el subíndice j=1,2……..g
  - **"p"** variables
  - **"n"** elementos para cada una de ellas que representamos con i=1,2…….n
- **Cada observación de cada variable para cada sujeto la representaremos como:**



*Así, el término X_{123} representaría la observación de la **variable 2** para el **elemento 1** del **grupo 3***

ANÁLISIS DISCRIMINANTE

SELECCIÓN DEL MÉTODO (II) *(Continuación)*

(Obtención de la/s funciones discriminantes) *(Continuación)*

- **Conforme a esta nomenclatura definimos las siguientes matrices:**

MATRIZ DE OBSERVACIONES PARA EL ELEMENTO "i" DEL GRUPO "j"

$$X_{ij} = \begin{pmatrix} X_{i1j} \\ X_{i2j} \\ . \\ . \\ X_{ipj} \end{pmatrix} \forall \begin{matrix} i = 1,2,....n_j \\ j = 1,2.....g \end{matrix}$$

MATRIZ DE MEDIAS DEL GRUPO "j"

$$\overline{X}_{\bullet j} = \begin{pmatrix} \overline{X}_{\bullet 1j} \\ \overline{X}_{\bullet 2j} \\ . \\ . \\ \overline{X}_{\bullet pj} \end{pmatrix} \forall \; j = 1,2.....g$$

MATRIZ DE MEDIAS TOTALES

$$\overline{X} = \begin{pmatrix} \overline{X}_{\bullet 1\bullet} \\ \overline{X}_{\bullet 2\bullet} \\ . \\ . \\ \overline{X}_{\bullet p\bullet} \end{pmatrix}$$

ANÁLISIS DISCRIMINANTE

SELECCIÓN DEL MÉTODO (II) *(Continuación)*

(Obtención de la/s funciones discriminantes) *(Continuación)*

- **Definidas estas matrices la variación Entre e Intra será:**

$$E = \sum_{j=1}^{g} n_j \cdot (\bar{X}_{\bullet j} - \bar{X})(\bar{X}_{\bullet j} - \bar{X})'$$

[matriz de orden (p x p)]

$$I = \sum_{j=1}^{g} \sum_{i=1}^{n_j} \cdot (X_{ij} - \bar{X}_{\bullet j})(X_{ij} - \bar{X}_{\bullet j})'$$

[matriz de orden (p x p)]

- **De esta forma, la ratio a maximizar sería:**

$$F = \frac{V.Entre}{V.Intra}$$

- Sin embargo nuestro **objetivo es encontrar los parámetros "b"** de la combinación lineal:

$$Y = b' X$$

que maximicen este ratio por lo que debemos expresar estas V.Intra y V.Entre en función de los parámetros “b” de este modelo.

ANÁLISIS DISCRIMINANTE

SELECCIÓN DEL MÉTODO (II) *(Continuación)*

(Obtención de la/s funciones discriminantes) *(Continuación)*

- **Dicho de otro modo, lo que queremos es maximizar la V. Entre y minimizar la V.Intra para la variable discriminante "y". Puede demostrarse que:**

$$SCE_y = b'Eb$$

$$SCI_y = b'I\ b$$

- **Por lo que, lógicamente, el ratio a maximizar puede expresarse como:**

$$F = \frac{SCE_y / g - 1}{SCI_y / n - g}$$

que obviando los grados de libertad supone:

$$max(l) = max \frac{b' Eb}{b' Ib}$$

- Esta operación arroja varias soluciones del conjunto de parámetros "b" lo que significa que para un determinado conjunto de datos siempre encontraremos más de una solución. (El menor de "g-1" o "p").

ANÁLISIS DISCRIMINANTE

F- VALIDACIÓN DE LAS FUNCIONES DISCRIMINANTES

- Autovalores: En el método de Fisher, la obtención de las distintas funciones se deriva de un proceso de obtención de raíces y vectores propios de una forma cuadrática. La suma de cuadrados entre grupos de cada función discriminante, viene definida por un autovalor $\lambda(i)$.

- Ratio Autovalor / Suma autovalores: capacidad discrimnante relativa, pero no absoluta.

- Test Bartlett: El test de Bartlett, distribuido como una χ^2 con p(g-1) grados de libertad, contrasta "secuencialmente" la hipótesis $(H_0 : l_1 = l_2 = ... = l_r = 0)$ presenta la siguiente forma:

$$B = \left[n - 1 - \frac{p+g}{2} \right] \sum_{j=1}^{r} 1n(1 + l_j)$$

- Correlación canónica función - variable clasificación original: Coeficientes elevados anticipan adecuada capacidad discriminante

G.- CONTRIBUCIÓN DE LAS VARIABLES A LA CAPACIDAD DISCRIMINANTE DE LAS FUNCIONES

- ANOVA **simple** con cada variable y la agrupación previa

- Parámetros estandarizados de la(s) función(es) discriminantes

- "CARGAS" DISCRIMINANTES: correlaciones entre cada variable inicial "x" y las funciones discriminantes "y".

ANÁLISIS DISCRIMINANTE

H.- VALORACIÓN DE LA CAPACIDAD PREDICTIVA

- Los contrastes de significación no informan sobre la capacidad predictiva del modelo
- Cálculo de la Puntuación de Corte Óptima
- Cálculo de la Puntuación de Corte Óptima modificada para el caso de grupos de tamaño desigual representativos de la estructura de la población (*muestreo aleatorio*).
- Construcción de la "Matriz de Confusión"
- Análisis de casos individuales (detección de nuevas variables a incluir en el análisis)

I.- UTILIZACIÓN FUNCIONES

- **Cálculo de la puntuación discriminante** (*estandarizadas – útil para la interpretación o no estandarizadas – útiles para el cálculo final*)
- **Cálculo de las Funciones Discriminantes Lineales de Fisher o Funciones de Clasificación** (*una por grupo, interesantes para simplificar la clasificación de nuevos elementos: clasificación en el grupo de mayor valor para su FDLF*)
- **Cálculo de los Centroides** y contraste de diferencias significativas aún en el caso de que se las funciones sean plenamente significativas (*D de Mahalanobis)*
- **Dibujo de los Centriodes** y distribuciones alrededor de los mismos para 2 ó 3 (*max*) funciones discriminantes