4. DISTANCIAS MULTIVARIADAS. MEDIDAS Y PRUEBAS

Distancias multivariadas

Numerosos problemas multivariados pueden analizarse en términos de ***distancias*** entre:

- Observaciones
- Muestras de observaciones
- Poblaciones de observaciones

Ej. 3 Distribución de una mariposa

4 variables ambientales

6 frecuencias genéticas

16 colonias de mariposas

Distancias (ambientales y genéticas)

¿Existe alguna relación entre ambos conjuntos de distancias?

Distancias multivariadas

- Se han propuesto diversos **métodos para medir** estas **distancias** y utilizarlas en el análisis multivariado.

A continuación analizaremos los métodos más comunes.

Una cierta cuota de arbitrariedad parece inevitable

Distancias entre observaciones individuales

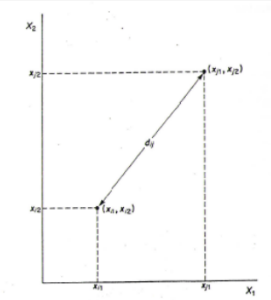
Sean p variables X_1, X_2, ….., X_p y n observaciones

$$\begin{array}{c} \\ 1 \\ \\ i \\ j \\ \\ n \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & x_{1p} \\ \\ x_{i1} & x_{i2} & x_{ip} \\ x_{j1} & x_{j2} & x_{jp} \\ \\ x_{n1} & x_{n2} & x_{np} \end{bmatrix} \end{array}$$

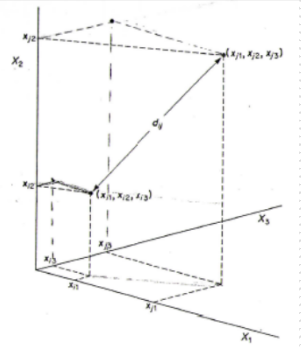
¿Cómo medir la "distancia" entre las observaciones ***i*** y ***j*** ?

Distancias entre observaciones individuales

i) p = 2 variables



ii) p = 3 variables



Distancia euclídea (Pitágoras):

$$d_{ij} = \sqrt{\sum_{k=1}^{2} (x_{ik} - x_{jk})^2}$$

$$d_{ij} = \sqrt{\sum_{k=1}^{3} (x_{ik} - x_{jk})^2}$$

Distancias entre observaciones individuales

iii) p variables (espacio p-dimensional)

generalización de la distancia euclidiana: $d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$

Si una de las variables observadas es mucho más variable que las otras, ésta predominará en el cálculo de las distancias **(Problema de Escala)**

En términos prácticos, es deseable que todas las variables tengan igual peso en el cálculo de las distancias. Esto se logra estandarizando previamente las variables originales.

Distancias entre observaciones individuales

Ej. 4.1 Distancias entre perros y especies relacionadas.

Los datos de la Tabla 1.4 muestran medidas medias de la mandíbula de 7 grupos de perros y especies relacionadas. Recordemos que la interrogante principal es: ¿Cómo se relacionan los perros prehistóricos con los otros grupos?

Distancias entre observaciones individuales

Cálculo de las distancias:

1. **Estandarizar las medidas.** Por ejemplo, expresándolas como desviaciones de las medias en unidades de desviación estándar.

X_1	X_2	X_3	X_4	X_5	X_6
-0,46	-0,46	-0,68	-0,69	-0,45	-0,57
-1,41	-1,79	-1,04	-1,29	-0,80	-1,21
1,78	1,48	1,70	1,80	1,55	1,50
0,60	0,55	0,96	0,69	1,17	0,88
0,13	0,31	-0,04	0,00	-1,10	-0,37
-0,52	0,03	-0,13	-0,17	0,03	0,61
-0,11	-0,12	-0,78	-0,34	-0,41	-0,83

Distancias entre observaciones individuales

2. Cálculo de las distancias euclidianas entre los 7 grupos

	Modern dog	Golden jackal	Chinese wolf	Indian wolf	Cuon	Dingo	Prehistoric dog
Modern dog	---						
Golden jackal	1,91	---					
Chinese wolf	5,38	7,12	---				
Indian wolf	3,38	5,06	2,14	---			
Cuon	1,51	3,19	4,57	2,91	---		
Dingo	1,56	3,18	4,21	2,20	1,67	---	
Prehist dog	**0,66**	2,39	5,12	3,24	1,26	1,71	---

El perro prehistórico es bastante similar al perro moderno de Tahilandia, puesto que la distancia entre esos dos grupos es la menor en toda la tabla.

Distancias entre poblaciones y muestras

Existen numerosas propuestas para determinar la **distancia entre dos poblaciones multivariadas**, siempre que las medias, varianzas y covarianzas de las poblaciones sean conocidas.

- **Penrose, 1953**
- **Mahalanobis, 1948**

Distancias entre poblaciones y muestras

- **Penrose, 1953**

  Supongamos que se disponen ***2 o más poblaciones*** y que las distribuciones en esas poblaciones se conocen para ***p variables*** X_1, X_2,……,X_p.

  Sea μ_{ki} la media de la variable X_k en la *i*-ésima población, y asumamos que la varianza de X_k tiene el mismo valor, V_k, en todas poblaciones.

$$P_{ij} = \sum_{k=1}^{p} \frac{(\mu_{ki} - \mu_{kj})^2}{p\, V_k}$$

representa la **distancia entre la población *i* y la población *j***

Distancias entre poblaciones y muestras

Desventaja

La distancia de Penrose no toma en cuenta las correlaciones entre las p variables. Esto significa que, cuando dos variables representan esencialmente lo mismo, es decir, están fuertemente correlacionadas, aun así ambas contribuyen de igual modo en el cálculo de las distancias entre las poblaciones, exactamente del mismo modo que lo hace una tercer variable totalmente independiente de las restantes variables.

Distancias entre poblaciones y muestras

- **Mahalanobis, 1948**
  Distancia entre la población *i* y la población *j*

  Esta es una medida que toma en cuenta la correlación entre las variables

$$D_{ij}^2 = \sum_{r=1}^{p}\sum_{s=1}^{p}\left(\mu_{ri} - \mu_{rj}\right)v^{rs}\left(\mu_{si} - \mu_{sj}\right)$$

  donde v^{rs} es el elemento ubicado en la fila r y columna s de la matriz inversa de la matriz de covarianza para nuestras p variables.

Distancias entre poblaciones y muestras

En forma vectorial: $D_{ij}^2 = (\boldsymbol{\mu_i} - \boldsymbol{\mu_j})' \mathbf{V}^{-1} (\boldsymbol{\mu_i} - \boldsymbol{\mu_j})$

donde: $\boldsymbol{\mu_i} = \begin{bmatrix} \mu_{1i} \\ \mu_{2i} \\ . \\ . \\ \mu_{pi} \end{bmatrix}$ es el vector de medias de la población i,

y **V** es la matriz de covarianza.

Obs.! Esta medida solamente puede ser calculada si la matriz de covarianza es la misma para todas las poblaciones.

Distancias entre poblaciones y muestras

- **Distancia de una observación respecto al centro de la población de donde proviene.**

  En este caso también se utiliza la **distancia de Mahalanobis.**

  Si x_1, x_2, ….,x_p representan los valores de una observación de las variables X_1, X_2, …, X_p, y μ_1, μ_2,…, μ_p representan las medias de la población, luego:

$$D^2 = \sum_{r=1}^{p}\sum_{s=1}^{p}(x_r - \mu_r)v^{rs}(x_s - \mu_s) = (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{V^{-1}} (\boldsymbol{x} - \boldsymbol{\mu})$$ donde:

Elemento en la r-ésima fila y s-ésima columna de la matriz inversa de V

$$\boldsymbol{x'} = (x_1, x_2, ..., x_p)$$

$$\boldsymbol{\mu'} = (\mu_1, \mu_2, ..., \mu_p)$$

Distancias entre poblaciones y muestras

V = matriz de covarianza de la población

$\mathbf{v^{rs}}$ = elemento en la fila r y columna s de la matriz $\mathbf{V^{-1}}$

D^2 representa un residual multivariado para la observación **x**. El término "residual" representa aquí una medida de cuán lejos se encuentra la observación **x** del centro de las distribuciones de todos los valores, tomando en cuenta todas las variables consideradas y sus covarianzas.

Distancias entre poblaciones y muestras

Propiedad

Si la población considerada tiene una distribución multinormal, los valores de D^2 siguen una distribución X^2 con **p** g.d.l si **x** proviene de esa distribución.

Un **valor elevado de D^2** significa que la correspondiente observación es

(a) un dato genuino y anómalo

(b) una observación de otra distribución

(b) un dato erróneo

Surge, por lo tanto, la necesidad de verificar si la observación es correcta o errónea.

Distancias entre poblaciones y muestras

Las ecuaciones presentadas en este subcapítulo presuponen que se utilizan estimaciones de la media, varianza y covarianza de la población en sustitución de los verdaderos valores.

En este caso, la matriz de covarianza **V** se sustituye por la estimación en base a todas las muestras disponibles:

$$C = \sum_{i=1}^{m} (n_i - 1)\, C_i \,/ \sum_{i=1}^{m} (n_i - 1)$$

m= nro. de muestras, $\mathbf{n_i}$= tamaño de la i-ésima muestra

$\mathbf{C_i}$= matriz de covarianza de la i-ésima muestra, con n-1 g.d.l.

C= estimación de la matriz de covarianza, con $\sum_{i=1}^{m} (n_i - 1)$ g.d.l.

Distancias entre poblaciones y muestras

- En principio, la distancia de Mahalanobis > distancia de Penrose, puesto que utiliza información de la covarianza.
- Esta ventaja es válida si las covarianzas se conocen con precisión. En cambio, si las covarianzas solo pueden ser estimadas con un pequeño número de muestras, probablemente sea mejor Penrose.
- ¿Qué se entiende por "pequeño número de muestras"? No es fácil precisarlo, en general no hay problema en utilizar la distancia de Mahalanobis con base en una matriz de covarianza estimada a partir de 100 o más muestras.

Distancias entre poblaciones y muestras

- **Ej. 4.2 Distancias entre muestras de cráneos egipcios**

Las **5 muestras** dan lugar a los siguientes vectores de medias:

$$\bar{x}_1 = \begin{bmatrix} 131{,}37 \\ 133{,}60 \\ 99{,}17 \\ 50{,}53 \end{bmatrix} \quad \bar{x}_2 = \begin{bmatrix} 132{,}37 \\ 132{,}70 \\ 99{,}07 \\ 50{,}23 \end{bmatrix} \quad \bar{x}_3 = \begin{bmatrix} 134{,}47 \\ 133{,}80 \\ 96{,}03 \\ 50{,}57 \end{bmatrix}$$

$$\bar{x}_4 = \begin{bmatrix} 135{,}50 \\ 132{,}30 \\ 94{,}53 \\ 51{,}97 \end{bmatrix} \quad \bar{x}_5 = \begin{bmatrix} 136{,}17 \\ 130{,}33 \\ 93{,}50 \\ 51{,}37 \end{bmatrix}$$

4 variables medidas

15 observaciones

Distancias entre poblaciones y muestras

y matrices de covarianza:

$$C_1 = \begin{bmatrix} 26{,}31 & 4{,}15 & 0{,}45 & 7{,}25 \\ 4{,}15 & 19{,}97 & -0{,}79 & 0{,}39 \\ 0{,}45 & -0{,}79 & 34{,}63 & -1{,}92 \\ 7{,}25 & 0{,}39 & -1{,}92 & 7{,}64 \end{bmatrix} \quad C_2 = \begin{bmatrix} 23{,}14 & 1{,}01 & 4{,}77 & 1{,}84 \\ 1{,}01 & 21{,}60 & 3{,}37 & 5{,}62 \\ 4{,}77 & 3{,}37 & 18{,}89 & 0{,}19 \\ 1{,}84 & 5{,}62 & 0{,}19 & 8{,}74 \end{bmatrix}$$

$$C_3 = \begin{bmatrix} 12{,}12 & 0{,}79 & -0{,}78 & 0{,}90 \\ 0{,}79 & 24{,}79 & 3{,}59 & -0{,}09 \\ -0{,}78 & 3{,}59 & 20{,}72 & 1{,}67 \\ 0{,}90 & -0{,}09 & 1{,}67 & 12{,}60 \end{bmatrix} \quad C_4 = \begin{bmatrix} 15{,}36 & -5{,}53 & -2{,}17 & 2{,}05 \\ -5{,}53 & 26{,}36 & 8{,}11 & 6{,}15 \\ -2{,}17 & 8{,}11 & 21{,}09 & 5{,}33 \\ 2{,}05 & 6{,}15 & 5{,}33 & 7{,}96 \end{bmatrix}$$

$$C_5 = \begin{bmatrix} 28{,}63 & -0{,}23 & -1{,}88 & -1{,}99 \\ -0{,}23 & 24{,}71 & 11{,}72 & 2{,}15 \\ -1{,}88 & 11{,}72 & 25{,}57 & 0{,}40 \\ -1{,}99 & 2{,}15 & 0{,}40 & 13{,}83 \end{bmatrix}$$

Distancias entre poblaciones y muestras

- Las matrices de covarianza parecen diferir. Sin embargo, recordemos que en la sección anterior (Ejemplo 3.3) se demostró que las diferencias no son significativas. Por tanto, parece razonable utilizar una matriz media ponderada:

$$C = \begin{bmatrix} 21{,}111 & 0{,}038 & 0{,}078 & 2{,}010 \\ 0{,}038 & 23{,}486 & 5{,}200 & 2{,}844 \\ 0{,}078 & 5{,}200 & 24{,}180 & 1{,}134 \\ 2{,}010 & 2{,}844 & 1{,}134 & 10{,}154 \end{bmatrix}$$

Distancias entre poblaciones y muestras

- **Distancias de Penrose** entre cada par de muestras:

p = 4 variables. Las varianzas estimadas (matriz C) son:

$$\hat{V}_1 = 21{,}112\,;\ \hat{V}_2 = 23{,}486\,;\ \hat{V}_3 = 24{,}180\,;\ \hat{V}_4 = 10{,}154$$

La distancia entre la muestra 1 y la muestra 2 es:

$$P_{ij} = \sum_{k=1}^{p} \frac{(\mu_{ki} - \mu_{kj})^2}{p\,V_k}$$

$$P_{12} = \frac{(131{,}37 - 132{,}37)^2}{4 \times 21{,}112} + \frac{(133{,}60 - 132{,}70)^2}{4 \times 23{,}486} + \frac{(99{,}17 - 99{,}07)^2}{4 \times 24{,}180} + \frac{(50{,}53 - 50{,}23)^2}{4 \times 10{,}154} = 0{,}023$$

Tiene significado si se compara con otro par de muestras

Distancias entre poblaciones y muestras

Procediendo de la misma forma, se obtiene la matriz de distancias:

Predinastía temprana	–				
Predinastía tardía	0,023	–			
12-13 dinastías	0,216	0,163	–		
Ptolomeica	0,493	0,404	0,108	–	
Romana	0,736	0,583	0,244	0,066	–

Recordemos del Ej. 3.3 que los valores medios cambian significativamente de muestra en muestra. Las distancias de Penrose muestran que **los cambios son acumulativos en el tiempo: las muestras cercanas en el tiempo son relativamente similares mientras que muestras lejanas en el tiempo son muy diferentes.**

Distancias entre poblaciones y muestras

Distancias de Mahalanobis

La inversa de la matriz C es:

$$C^{-1} = \begin{bmatrix} 0{,}0483 & 0{,}0011 & 0{,}0001 & -0{,}0001 \\ 0{,}0011 & 0{,}0461 & -0{,}0094 & -0{,}0121 \\ 0{,}0001 & -0{,}0094 & 0{,}0435 & -0{,}0022 \\ -0{,}0099 & -0{,}0121 & -0{,}0022 & 0{,}1041 \end{bmatrix}$$

y la distancia entre la muestra 1 y 2:

$$D_{ij}^2 = \sum_{r=1}^{p}\sum_{s=1}^{p}\left(\mu_{ri} - \mu_{rj}\right) v^{rs} \left(\mu_{si} - \mu_{sj}\right)$$

$$D_{12}^2 = (131{,}37 - 132{,}37)0{,}0483(131{,}37 - 132{,}37) + (131{,}37 - 132{,}37)0{,}0011(133{,}60 - 132{,}70) + \cdots$$
$$- (50{,}53 - 50{,}23)0{,}0022(99{,}17 - 99{,}07) + (50{,}53 - 50{,}23)0{,}1041(50{,}53 - 50{,}23) = 0{,}091$$

Distancias entre poblaciones y muestras

Procediendo de la misma forma, se obtiene la matriz de distancias:

Predinastía temprana	–				
Predinastía tardía	0,091	–			
12-13 dinastías	0,903	0,729	–		
Ptolomeica	1,881	1,594	0,443	–	
Romana	2,697	2,176	0,911	0,219	–

Distancias entre poblaciones y muestras

Una comparación entre las distancias de Penrose y Mahalanobis muestran buena concordancia, siendo estas últimas 3 a 4 veces mayores que las distancias de Penrose.

Penrose

$$\begin{bmatrix} - & & & & \\ 0{,}023 & - & & & \\ 0{,}216 & 0{,}163 & - & & \\ 0{,}493 & 0{,}404 & 0{,}108 & - & \\ 0{,}736 & 0{,}583 & 0{,}244 & 0{,}066 & - \end{bmatrix}$$

Mahalanobis

$$\begin{bmatrix} - & & & & \\ 0{,}091 & - & & & \\ 0{,}903 & 0{,}729 & - & & \\ 1{,}881 & 1{,}594 & 0{,}443 & - & \\ 2{,}697 & 2{,}176 & 0{,}911 & 0{,}219 & - \end{bmatrix}$$

$\frac{0{,}736}{0{,}023} = 32{,}0$ **La distancia relativa entre muestras es prácticamente similar para ambas medidas** $\frac{2{,}697}{0{,}091} = 29{,}6$

Distancias en base a proporciones

- **Caso particular:** Las variables están expresadas como proporciones, cuya suma es la unidad.

Ejemplo: Los animales de cierta especie pueden clasificarse en K clases genéticas:

Colonia 1: $p_1, p_2, \ldots\ldots, p_k$ proporción p_1 la clase 1, etc.

Colonia 2: $q_1, q_2, \ldots\ldots, q_k$ proporción q_1 la clase 1, etc.

¿Cuán similares son las colonias 1 y 2 en términos genéticos?

Distancias en base a proporciones

Índices de distancia (para datos expresados como proporciones):

1) $d_1 = \sum_{i=1}^{k} |p_i - q_i| / 2$

toma el valor 1 cuando no existe solapamiento entre clases y el valor 0 cuando $p_i = q_i$ para todo i.

2) $d_2 = 1 - \sum_{i=1}^{k} p_i q_i \Big/ \sqrt{\sum_{i=1}^{k} p_i^2 \sum_{i=1}^{k} q_i^2}$

varía desde 1(no solapamiento) a 0 (iguales proporciones)

Distancias en base a proporciones

Debido a que d_1 y d_2 varían de 0 a 1, se sigue que $1-d_1$ y $1-d_2$ son medidas de la similitud entre los casos que se comparan. De hecho, es en términos de similitudes que los índices se utilizan a menudo. Por ejemplo,

$$s_1 = 1 - d_1 = 1 - \sum_{i=1}^{k} |p_i - q_i| / 2$$

se utiliza a menudo como una medida de la superposición de nichos entre dos especies, donde p_i es la fracción de los recursos utilizados por las especies 1 que son de tipo i y q_i es la fracción de los recursos utilizados por las especies 2 que son de tipo i. Entonces **$s_1 = 0$ indica que las dos especies usan recursos completamente diferentes**, y **$s_1 = 1$ indica que las dos especies utilizan exactamente los mismos recursos.**

Presencia y ausencia de datos

Presencia y Ausencia de 2 especies en 10 sitios

Sitio	1	2	3	4	5	6	7	8	9	10
Especie 1	0	0	1	1	1	0	1	1	1	0
Especie 2	1	1	1	1	0	0	0	0	1	1

1 = presencia
2 = ausencia

La similitud o distancia entre dos ítems debe basarse en una lista de sus presencias y ausencias. Por ejemplo, podría haber interés en la **similitud entre 2 especies de plantas en términos de sus distribuciones en 10 sitios.**

(a) ambas especies están presentes a=3
(b y c) sólo una especie está presente b=3 c=3
(d) ambas especies están ausentes d=1

Presencia y ausencia de datos

Observación:

Estos índices varían de 0 (sin similitud) a 1 (semejanza completa). Algunos autores revisaron estos índices con base en las presencias y ausencias de 25 especies de especies de peces en 52 lagos y han debatido si el número de ausencias conjuntas (d) debe utilizarse en el cálculo, debido al peligro de concluir que dos especies son similares simplemente porque están ausentes de muchos sitios. Este es sin duda un punto válido en muchas situaciones, y **sugiere que el índice de concordancia simple debe utilizarse con precaución.**

Prueba de aleatorización de Mantel (1967)

Objetivo: Comparar dos matrices de distancias o de similaridad.

Supongamos que se estudian 4 objetos y que, para cada uno de ellos,se han medido 2 conjuntos de variables.

Con el primer conjunto de variables puede construirse una matriz 4 x 4:

$$M = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} = \begin{bmatrix} 0.0 & 1.0 & 1.4 & 0.9 \\ 1.0 & 0.0 & 1.1 & 1.6 \\ 1.4 & 1.1 & 0.0 & 0.7 \\ 0.9 & 1.6 & 0.7 & 0.0 \end{bmatrix}$$

donde m_{ij} = distancia entre los objetos ***i*** y ***j***.

Prueba de aleatorización de Mantel (1967)

i) M es simétrica

ii) Los elementos de la diagonal son ceros puesto que representan la distancia de los objetos consigo mismo.

Con el segundo conjunto de variables puede también construirse una matriz de distancias:

$$E = \begin{bmatrix} e_{11} & e_{12} & e_{13} & e_{14} \\ e_{21} & e_{22} & e_{23} & e_{24} \\ e_{31} & e_{32} & e_{33} & e_{34} \\ e_{41} & e_{42} & e_{43} & e_{44} \end{bmatrix} = \begin{bmatrix} 0.0 & 0.5 & 0.8 & 0.6 \\ 0.5 & 0.0 & 0.5 & 0.9 \\ 0.8 & 0.5 & 0.0 & 0.4 \\ 0.6 & 0.9 & 0.4 & 0.0 \end{bmatrix}$$

Prueba de aleatorización de Mantel (1967)

La prueba/test de Mantel tiene por objetivo evaluar si los elementos en las matrices **M** y **E** están correlacionados.

Para matrices n x n se cálcula el estadígrafo: $Z = \sum_{i=2}^{n} \sum_{j=1}^{i-1} m_{ij} e_{ij}$

que representa la suma de los productos de los elementos debajo de la diagonal de las matrices **M** y **E**.

El estadígrafo Z se compara con la distribución de Z que se obtiene **tomando los objetos en orden aleatorio para una de las dos matrices.** De ahí "prueba de aleatorización".

Prueba de aleatorización de Mantel (1967)

Ejemplo: M puede permanecer sin modificaciones. Para los objetos de la matriz E puede seleccionarse un orden aleatorio: 3, 2, 4, 1. Con lo que se obtiene la matriz aleatoria:

$$E_R = \begin{bmatrix} e_{33} & e_{32} & e_{34} & e_{31} \\ e_{23} & e_{22} & e_{24} & e_{21} \\ e_{43} & e_{42} & e_{44} & e_{41} \\ e_{13} & e_{12} & e_{14} & e_{11} \end{bmatrix} = \begin{bmatrix} 0.0 & 0.5 & 0.4 & 0.8 \\ 0.5 & 0.0 & 0.9 & 0.5 \\ 0.4 & 0.9 & 0.0 & 0.6 \\ 0.8 & 0.5 & 0.6 & 0.0 \end{bmatrix} \begin{matrix} 3 \\ 2 \\ 4 \\ 1 \end{matrix}$$

Un valor del estadígrafo Z puede calcularse utilizando M y E_R.

Prueba de aleatorización de Mantel (1967)

Este procedimiento se puede repetir utilizando diferentes ordenes aleatorios para generar $\mathbf{E_R}$ y la distribución aleatoria Z.

La prueba consiste en verificar si el valor observado de Z es un valor típico de esta distribución.

Idea básica

Si dos medidas de distancia están completamente no relacionadas, la matriz **E** debe ser justamente igual a una de las matrices $\mathbf{E_R}$, ordenadas aleatoriamente. De aquí, el valor observado Z será un valor típicamente aleatorio de Z.

Prueba de aleatorización de Mantel (1967)

Por otra parte, si las dos distancias tienen una correlación positiva, luego el valor observado de Z tenderá a ser mayor que los valores Z resultantes de aplicar el ordenamiento aleatorio.

- **n objetos** → **n! diferentes ordenes posibles**
- **n! matrices $\mathbf{E_R}$**, algunas de las cuales pueden generar idénticos valores de Z.

En nuestro ejemplo n=4 y n!=24 valores de Z. Pero si n=5, n!=1,3 x 10^{12}. Esto puede resultar poco práctico, con lo que se plantean dos alternativas para efectuar la prueba/test de Mantel:

Prueba de aleatorización de Mantel (1967)

1) Generar computacionalmente un número importante de matrices E_R y tomar la distribución resultante de los valores Z como verdadera (lo sería si tomamos todas las posibles matrices E_R).

2) Calcular la media {E(Z)} y la varianza {var(Z)} de la distribución aleatoria de Z , y

$$g = \frac{Z - E(Z)}{\sqrt{Var(Z)}}$$

puede tratarse como normal estándar.

Prueba de aleatorización de Mantel (1967)

Mantel propuso ecuaciones para calcular E(Z) y var(Z) en el caso de la hipótesis nula (no correlación entre las distancias). No obstante, existen dudas respecto a la validez de la aproximación normal del estadígrafo g (Mielke, 1978).

Parece mejor utilizar la alternativa 1.

Prueba de aleatorización de Mantel (1967)

Los valores de Z no son particularmente informativos, excepto en comparación con la media y la varianza. Por lo tanto, puede resultar más útil tomar la correlación entre los elementos debajo de la diagonal de las matrices M y E como estadígrafo, en lugar de Z.

$$r_{ME} = \frac{Z - n(n-1)\overline{m}\,\overline{e}\,/\,2}{\sqrt{\left(\sum_{i=2}^{n}\sum_{j=1}^{i-1} m_{ij}^2 - n(n-1)\overline{m}^2/2\right)\left(\sum_{i=2}^{n}\sum_{j=1}^{i-1} e_{ij}^2 - n(n-1)\overline{e}^2/2\right)}}$$

donde:

$\overline{m} = E(m_{ij})$

$\overline{e} = E(e_{ij})$

$\frac{n(n-1)}{2} =$ Nro. de elementos debajo de la diagonal de las matrices M y E.

Prueba de aleatorización de Mantel (1967)

Ventaja:

-1 < r < +1

r = - 1 correlación negativa

r = 0 no correlación

r = +1 correlación positiva

La significancia de los datos es la misma para el estadígrafo Z y r, puesto que r es justamente Z menos una constante, dividida por otra constante.

Prueba de aleatorización de Mantel (1967)

Ej. 4.3 Más sobre distancias entre muestras de cráneos egipcios

Las distancias obtenidas en el Ej. 4.2, en base a 4 medidas de los cráneos, ¿son significativas en relación a las diferencias temporales entre las 5 muestras?

Ese parece ser el caso. Apliquemos la prueba de Mantel para responder categóricamente a esta interrogante.

Prueba de aleatorización de Mantel (1967)

Los tiempos de las muestras son aproximadamente:

- período predinastía temprana (4000 AC)
- periodo predinastía tardía (3300 AC)
- 12 y 13 dinastías (1850 AC)
- período Ptolemaico (200 AC)
- período romano (150 DC)

La comparación de las medidas de distancias de Penrose con las diferencias temporales (en miles de años) proporciona, por consiguiente, las siguientes matrices de distancias entre las muestras:

Prueba de aleatorización de Mantel (1967)

Distancias de Penrose

–				
0,023	–			
0,216	0,163	–		
0,493	0,404	0,108	–	
0,736	0,583	0,244	0,066	–

Distancias temporales

–				
0,70	–			
2,15	1,45	–		
3,80	3,10	1,65	–	
4,15	3,45	2,00	0,35	–

La correlación entre los elementos de esas matrices es **0,954**. Parece, por consiguiente, que las distancias concuerdan muy bien.

Prueba de aleatorización de Mantel (1967)

Existen 5! = 120 formas posibles de reordenar las cinco muestras para una de las dos matrices, y, en consecuencia, hay 120 elementos en la distribución aleatoria para la correlación. De éstos, uno es la correlación observada de 0,954 y otro es una correlación más grande. De ello se deduce que **la correlación observada es significativamente alta al nivel (2/120) 100% = 1,7%**, y hay evidencia de una relación entre las dos matrices de distancia.

La correlación matricial entre las **distancias de Mahalanobis** y las distancias temporales es **0,964**. Esto también es significativamente grande al nivel de 1,7% cuando se compara con la distribución de ramdomización.

Cómo realizar una una prueba de Mantel con XLSTAT?

- La prueba de Mantel se utiliza para calcular la correlación lineal entre dos matrices de proximidad (disimilaridad o similaridad).
- Una vez XLSTAT iniciado, elija el comando o barra de herramientas XLSTAT/Pruebas de correlación/Pruebas de Mantel.
- En el cuadro de diálogo seleccionar la primera matriz de disimilaridad (A), y luego la segunda (B). Activar las opciones "Distribución" y "Gráficos" con el fin de visualizar los resultados detallados. Presionar el botón "Más" para visualizar las opciones avanzadas. Si el tamaño de la matriz es menor de 10x10, seleccionar la opción "Exhaustivo" para que XLSTAT calcule todas las permutaciones posibles y el p-valor exacto. Activar la opción "Prueba bilateral" para que la hipótesis alternativa sea la hipótesis según la cual la correlación es diferente de cero.

Distancias Multivariadas - RESUMEN

- Muchos problemas multivariados se pueden considerar en términos de **distancias** entre pares de observaciones, muestras de observaciones o poblaciones de observaciones.
- La **distancia euclidiana** entre pares de observaciones individuales es una medida de distancia que se utiliza con frecuencia. Esto se puede ver con 2- o 3-variables como la distancia espacial entre las observaciones individuales cuando se representan gráficamente. Este concepto es generalizado para su uso con más de tres variables.
- Dos medidas de **distancia entre dos muestras multivariantes** o dos poblaciones multivariadas: la distancia de **Penrose**, que no toma en cuenta las correlaciones entre las variables, y la distancia de **Mahalanobis**, que toma en cuenta las correlaciones.

Distancias Multivariadas - RESUMEN

- Se describen **dos medidas de distancia** para la situación en la que **las p variables medidas sobre los objetos considerados son proporciones que se suman a uno**. Éstos se convierten en medidas de similitud rescatándolas de uno.

- Se describen **cuatro índices para medir la similitud** entre los objetos que se comparan, **sobre la base de la presencia y ausencia** de una serie de características.

- La **prueba de Mantel** se describe como un medio para determinar si dos matrices de distancias o similitudes muestran una asociación positiva o negativa significativa; y se utiliza para calcular la correlación lineal entre dos matrices de proximidad (disimilaridad o similaridad).