

Las técnicas del análisis multivariado tienen el objetivo de entender e interpretar datos que están medidos en espacios de gran dimensión, datos que son elementos de espacios multidimensionales.

Supóngase que tenemos un vector de variables \mathbf{x} en \mathbb{R}^p y sea x_1, x_2, \dots, x_n un conjunto de n observaciones de \mathbf{x} .

Tenemos

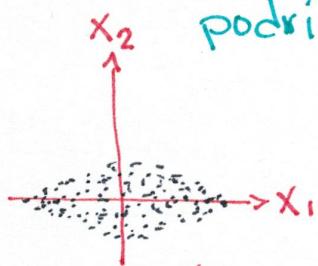
$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \quad y,$$

asimismo

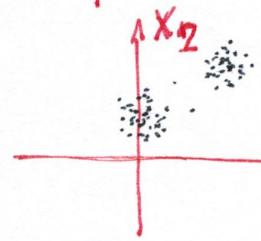
$$\mathbf{x} = (x_1, \dots, x_p)$$

Antes de pensar en llevar a cabo inferencias (estadísticas) a partir de x_1, \dots, x_n , deberíamos pensar en como describir, representar ó "ver" los datos. Si podemos visualizar ó representar gráficamente los datos, de forma que esto permita entender varias características importantes de los mismos, esto nos podría ayudar a seleccionar las herramientas matemáticas (estadísticas) adecuadas para hacer análisis e inferencias.

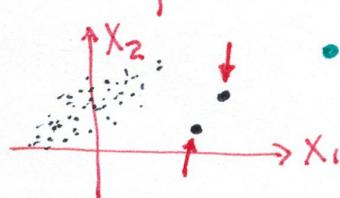
Algunas características de interés que los datos podrían presentar son:



- Algunos de los componentes de \mathbf{X} tienen una dispersión mayor a otras componentes.



- Hay componentes en \mathbf{X} que indican la existencia de subgrupos ó aglomeraciones en los datos.



- Existente "datos aberrantes" (outliers) en alguna componente de \mathbf{X} .



- La distribución "normal" multivariada no es un "buen" modelo para \mathbf{X} .

- Hay combinaciones lineales (de "baja dimensión") de los componentes de \mathbf{X} que tienen un comportamiento alejado del modelo "normal".

Los dibujos a la izquierda, nos permiten pensar que para $P=2$, los "diagramas de dispersión" son de ayuda para nuestros objetivos. Existen técnicas computacionales y programas de cómputo modernos, que permiten visualizar datos tridimensionales ($P=3$) así como rotaciones de los mismos en

tiempo real. ¿Qué técnicas podríamos usar si $p > 3$?

Revisaremos algunas técnicas de representación gráfica para datos multivariados cuya finalidad es describir características de los datos, como las mencionadas en la lista anterior.

DIAGRAMAS Ó GRÁFICAS DE CAJAS (BOX PLOTS)

Los diagramas de cajas son gráficos para representar la distribución de las variables X_1, X_2, \dots, X_p . Con estos se puede vislumbrar características de la distribución de cada componente X_i , las características podrían ser: localización, sesgo, dispersión, longitud (pesadez) de las colas y valores "aberrantes" (observaciones que aparecen fuera del rango de posibles valores de X_i).

En particular, las cajas permiten comparar estas características para dos (X_i y $X_j; i \neq j$) ó más variables (X_i, X_j y $X_k \quad i \neq j; j \neq k; i \neq k$)

Las cajas usan un resumen de los datos que consta de cinco estadísticas (Five-number summaries):

4

- El cuartil superior F_U
- El cuartil inferior F_L
- = La mediana
- = El mínimo y el máximo de la muestra
(los valores extremos de la muestra).

Consideremos una muestra de observaciones x_1, \dots, x_n y sean $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$ las correspondientes "estadísticas de orden" (los valores x_1, \dots, x_n ordenados de acuerdo a su magnitud). De esta forma $x_{(1)} = \min\{x_i : 1 \leq i \leq n\}$ y $x_{(n)} = \max\{x_i : 1 \leq i \leq n\}$

La mediana es el número que deja la mitad de los datos en $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$ antes de él y la mitad de los datos en $\{x_{(1)}, \dots, x_{(n)}\}$ después de él. La mediana se puede definir como

$$M \equiv \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ es impar,} \\ \frac{1}{2}\{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\} & \text{si } n \text{ es par.} \end{cases}$$

Para definir los cuartiles F_U y F_L :

1.- Usemos M para dividir $\{x_{(1)}, \dots, x_{(n)}\}$ en

dos subconjuntos (los $x_{(i)}$'s antes y después de M).

(a) Si hay un número impar de datos en $\{x_{(1)}, \dots, x_{(n)}\}$, incluimos M en las dos listas:

$$C_1 = \{x_{(1)}, x_{(2)}, \dots, x_{(\frac{n+1}{2})} = M\},$$

$$C_2 = \{x_{(\frac{n+1}{2})} = M, x_{(\frac{n+1}{2} + 1)}, \dots, x_{(n)}\}.$$

(b) Si hay un número par de datos en $\{x_{(1)}, \dots, x_{(n)}\}$, la mediana M no es un elemento a considerar, es decir las listas son:

$$C_1 = \{x_{(1)}, x_{(2)}, \dots, x_{(\frac{n}{2})}\} \quad y$$

$$C_2 = \{x_{(\frac{n}{2} + 1)}, x_{(\frac{n}{2} + 2)}, \dots, x_{(n)}\}$$

2: El cuartil F_L es la mediana de los datos C_1 .
El cuartil F_U es la mediana de los datos C_2 .

El cálculo de F_L y F_U usando los pasos 1 y 2 se conoce en literatura como "Método de Tukey"

ejemplo: Datos de tamaños poblacionales de las quince ciudades más grandes en 2006.

City	Country	Pop. (10,000)	Order statistics
Tokyo	Japan	3,420	$x_{(15)}$
Mexico city	Mexico	2,280	$x_{(14)}$
Seoul	South Korea	2,230	$x_{(13)}$
New York	USA	2,190	$x_{(12)}$
Sao Paulo	Brazil	2,020	$x_{(11)}$
Bombay	India	1,985	$x_{(10)}$
Delhi	India	1,970	$x_{(9)}$
Shanghai	China	1,815	$x_{(8)}$
Los Angeles	USA	1,800	$x_{(7)}$
Osaka	Japan	1,680	$x_{(6)}$
Jakarta	Indonesia	1,655	$x_{(5)}$
Calcutta	India	1,565	$x_{(4)}$
Cairo	Egypt	1,560	$x_{(3)}$
Manila	Philippines	1,495	$x_{(2)}$
Karachi	Pakistan	1,430	$x_{(1)}$

Para estos datos, $n=15$, $M=x_{(8)}=1.815$, $F_L=1.610$

$F_U=2.105$, $x_{(1)}=1.430$, $x_{(15)}=3.420$.

La F-dispersión d_F se define como

$$d_F \equiv F_U - F_L .$$

Las barres externas se definen como

$$b_U \equiv F_U + 1.5 \cdot d_F , \quad b_L \equiv F_L - 1.5 \cdot d_F$$

b_U y b_L son los números (los fronteras) más allá de los cuales, un dato se considera ó se clasifica como un valor aberrante (outlier).

Para los datos de tamaños poblacionales de las ciudades tenemos

$$d_F = F_U - F_L = 2105 - 1610 = 495$$

$$b_L = F_L - 1.5 \cdot d_F = 1610 - 1.5 \cdot 495 = 867.5$$

$$b_U = F_U + 1.5 \cdot d_F = 2105 + 1.5 \cdot 495 = 2847.5$$

El diagrama de caja también reporta la media de los datos

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 1939.7 \leftarrow \text{para los datos de tamaños poblacionales}$$

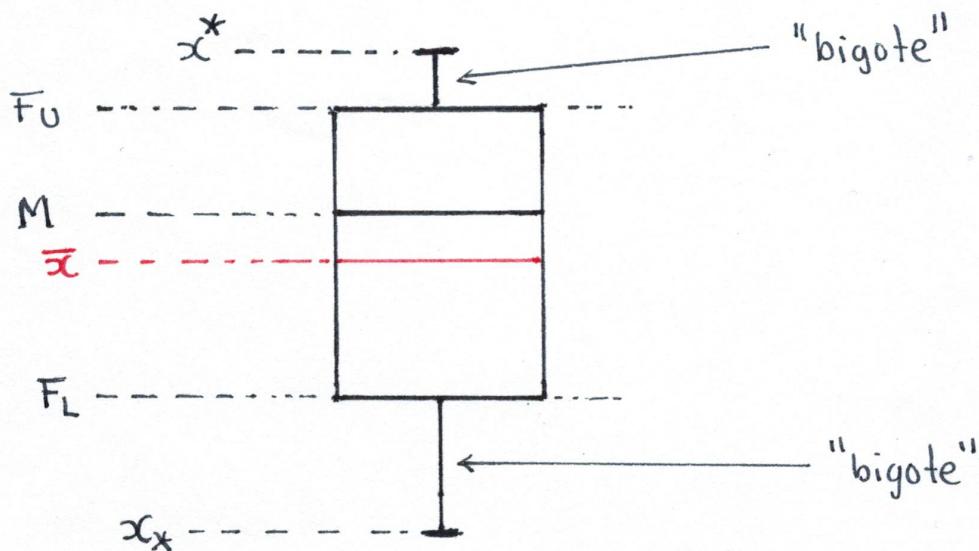
CONSTRUCCION DEL DIAGRAMA DE CAJA

- 1 Se dibuja una caja con bordes, arriba y abajo, en F_U y F_L respectivamente. El 50% de los datos quedan dentro de la caja.
- 2 Se dibuja una línea sólida en el valor de M y una línea punteada en el valor de \bar{x} .

3 Se dibujan líneas verticales ("bigotes") que van desde los bordes (arriba y abajo) de la caja y que llegan hasta los valores x^* y x_* dados por:

$$x^* = \max \{x \in \{x_{(1)}, \dots, x_{(n)}\} : x \leq b_U\},$$

$$x_* = \min \{x \in \{x_{(1)}, \dots, x_{(n)}\} : b_L \leq x\}.$$



4 Algunos paquetes de cómputo, muestran los datos aberrantes (outliers) con un carácter "*", si estos yacen fuera del intervalo (b_L, b_U) o con un carácter "●" si estos yacen fuera del intervalo $(F_L - 3 \cdot d_F, F_U + 3 \cdot d_F)$.

Para el ejemplo de los tamaños poblacionales de las ciudades tenemos que: