Tema 2: Estadística Descriptiva Multivariante

Datos multivariantes: estructura y notación

Se llama población a un conjunto de elementos bien definidos. Por ejemplo, la población de las empresas de un país, o de los estudiantes de una Universidad.

Cuando en cada elemento de la población se mide un conjunto de variables estadísticas diremos que se ha definido una variable estadística multivariante, vectorial o multidimensional. Las variables que se miden en cada elemento pueden ser cualitativas o cuantitativas. Algunos ejemplos de variables multivariantes son los siguientes:

(i) En cada estudiante de una universidad medimos la edad, el sexo, la nota de entrada en la universidad, el municipio de residencia y el curso más alto en que se encuentra matriculado.

(ii) En cada una de las empresas de un polígono industrial medimos el número de trabajadores, la facturación, el sector industrial y las ayudas oficiales recibidas.

(iii) En cada país del mundo medimos diez indicadores de desarrollo.

Supondremos en adelante que las variables definidas sobre cada elemento de la población son numéricas. En particular, cualquier variable cualitativa se transformará a una escala numérica. Por ejemplo, la variable sexo se convierte en numérica asignando el cero al varón y el uno a mujer. Naturalmente la asignación de valores numéricos es arbitraria. Entonces podemos suponer que los valores disponibles de la variable multidimensional

se encuentran en una matriz, que llamaremos matriz de datos. En esta matriz cada fila representa un elemento de la población y cada columna los valores de una variable escalar en todos los elementos observados. Típicamente esta matriz será rectangular con n filas y k columnas donde hemos supuesto que existen n elementos en la población y que se han medido k variables sobre cada elemento.

Llamaremos $\mathbf{X}$ a la matriz de datos y x_{ij} a su elemento genérico que representa el valor de la variable j sobre el individuo i. donde $i = 1, ..., n$ y $j = 1, ..., k$.

La matriz de datos $\mathbf{X}$ tendrá dimensiones $n \times k$ y puede representarse de dos formas distintas. Por filas como:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

donde cada variable $\mathbf{x}_i'$ es un vector fila $k \times 1$ que representa los valores de las k variables sobre el individuo i.

Alternativamente podemos representar la matriz $\mathbf{X}$ por columnas:

$$\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_k]$$

donde ahora cada variable $\mathbf{x}_i$ es un vector columna $n \times 1$ que representa la variable i, medida en los n elementos de la población.

Vector de Medias

La medida de centralización más utilizada para describir datos multivariantes es el vector de medias, que tiene dimensión k y recoge las medias de cada una de las k variables. Se calcula fácilmente mediante:

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_k \end{bmatrix} = \frac{1}{n}\mathbf{X}'\mathbf{1},$$

donde $\mathbf{1}$ representará siempre un vector de unos de la dimensión adecuada.

En R el comando correspondiente es:

```
mean(x)
```

y la cuasivarianza se calcula con el comando

```
var(x)
```

Ejemplo

La siguiente tabla presenta ocho variables físicas tomadas en un grupo de 27 estudiantes. Las variables son sexo (sex con 0 hombre, 1 mujer), estatura (*est*), peso en Kgr (*pes*), longitud de pie (*pie*), longitud de brazo (*lbr*), anchura de la espalda (*aes*), diámetro de cráneo (*dcr*) y longitud entre la rodilla y el tobillo (*drt*). Todas las longitudes van en cm (ver libro *Análisis Multivariante* de D. Peña (2001))

sex	*est*	*pes*	*pie*	*lbr*	*aes*	*dcr*	*drt*
0	159.0	49	36.0	68.0	42.0	57.0	40.0
1	164.0	62	39.0	73.0	44.0	55.0	44.0
0	172.0	65	38.0	75.0	48.0	58.0	44.0
...	...	...	...	...	...	...	...
0	170.0	70	38.0	73.0	45.0	56.0	43.0
1	170.0	67	40.0	77.0	46.5	58.0	44.5
0	168.0	56	37.5	70.5	48.0	60.0	40.0

La siguiente tabla presenta las medias y desviaciones típicas de las variables, así como otras medidas de la distribución univariante de cada variable.

	est	*pes*	*pie*	*lbr*	*aes*	*dcr*	*drt*
Medias	168.8	63.9	39.0	73.5	45.9	57.2	43.1
D. Típicas	10.2	12.8	2.9	4.9	4.0	1.8	3.1
Coef. asimetría	.15	.17	.27	.37	-.22	.16	.56
Coef. Curtosis	1.8	2.1	1.9	2.1	2.4	2.0	3.4

Se observa que la variable más homogénea (con menor variabilidad) es el diámetro del cráneo y la más variables el peso. La distribución más asimétrica es la distancia entre rodilla y tobillo y la más apuntada (con mayor curtosis) la distancia rodilla tobillo.

NOTA: En R se puede definir dos funciones para calcular la curtosis y la asimetría:

```
# funcion para calcular el coeficiente de asimetria de un vector de datos
asim <- function(x){
n <- length(x)
```

```
 asimetria <- (sum((x-mean(x))^3)/n) / ((sqrt(var(x))^3))
cbind(asimetria) }

# funcion para calcular el coeficiente de curtosis de un vector de datos
curto <- function(x){
n <- length(x)
kurtosis <- (sum((x-mean(x))^4)/n) / ((sqrt(var(x))^4)) - 3
cbind(kurtosis) }
```

Matriz de varianzas y covarianzas

La variabilidad de los datos y la información relativa a las relaciones lineales entre las variables se resumen en la matriz de varianzas y covarianzas. Esta matriz es cuadrada y simétrica de orden k, donde los términos diagonales son las varianzas y los no diagonales, las covarianzas entre las variables. Llamando $\mathbf{S}$ a esta matriz, tendremos que, por definición:

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_k^2 \end{bmatrix}.$$

Esta matriz puede calcularse como:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'.$$

La comprobación es inmediata. Como:

$$\begin{bmatrix} x_{i1} - \overline{x}_1 \\ \vdots \\ x_{ik} - \overline{x}_k \end{bmatrix} [x_{i1} - \overline{x}_1 \ldots \overline{x}_{ik} - \overline{x}_k] = \begin{bmatrix} (x_{i1} - \overline{x}_1)^2 & \cdots & (x_{i1} - \overline{x}_1)(x_{k1} - \overline{x}_k) \\ \vdots & \ddots & \vdots \\ (x_{ik} - \overline{x}_k)(x_{i1} - \overline{x}_1) & \cdots & (x_{ik} - \overline{x}_k)^2 \end{bmatrix}$$

al sumar para todos los elementos y dividir por n se obtienen las varianzas y covarianzas entre las variables. Otra forma de calcular $\mathbf{S}$ es a partir de la matriz de datos centrados $\widetilde{\mathbf{X}}$, que se obtiene restando a cada dato su media. Es fácil comprobar que esta matriz puede calcularse mediante

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}',$$

y sustituyendo el vector de medias por su expresión dada:

$$\widetilde{\mathbf{X}} = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X} = \mathbf{P}\mathbf{X},$$

donde la matriz cuadrada $\mathbf{P}$ está definida por

$$\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$$

y es simétrica e idempotente (ya que se puede comprobar que $\mathbf{P}\mathbf{P} = \mathbf{P}$). Entonces la matriz $\mathbf{S}$ puede escribirse:

$$\mathbf{S} = \frac{1}{n}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \frac{1}{n}\mathbf{X}'\mathbf{P}\mathbf{X}.$$

En R el comando correspondiente es:

```
cov(x)
```

Observación

La matriz de covarianzas es semidefinida positiva. Es decir, si $\mathbf{y}$ es cualquier vector $\mathbf{y}'\mathbf{S}\mathbf{y} \geq 0$.

Esta condición también implica que los autovalores de esta matriz λ_i son no negativos. Es decir, si $\mathbf{S}\mathbf{v}_i = \lambda_i \mathbf{v}_{i,}$, entonces $\lambda_i \geq 0$.

La matriz de correlación

Llamaremos matriz de correlación a la matriz cuadrada y simétrica que tiene unos en la diagonal y fuera de ella los coeficientes de correlación entre las variables. Escribiremos

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$

Esta matriz es también semidefinida positiva. Para verlo, llamemos $\mathbf{D}$ a la matriz diagonal de orden k construida colocando en la diagonal principal las desviaciones típicas de las variables. La matriz $\mathbf{R}$ esta relacionada con la matriz de covarianzas $\mathbf{S}$ mediante:

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1},$$

que implica

$$\mathbf{S} = \mathbf{DRD}.$$

La condición $\mathbf{w}'\mathbf{S}\mathbf{w} \geq 0$ equivale a:

$$\mathbf{w}'\mathbf{D}\mathbf{R}\mathbf{D}\mathbf{w} = \mathbf{Z}'\mathbf{R}\mathbf{Z} \geq \mathbf{0}$$

llamando $\mathbf{Z} = \mathbf{D}\mathbf{w}$. Por tanto, la matriz $\mathbf{R}$ es también semidefinida positiva.

En R el comando correspondiente es:

`corr(x)`

Correlaciones parciales

Se define la matriz de correlaciones parciales como la matriz que mide las relaciones entre pares de variables eliminando el efecto de las restantes. Por ejemplo, para cuatro variables:

$$\mathbf{R}_p = \begin{bmatrix} 1 & r_{12,34} & r_{13,24} & r_{14,23} \\ r_{12,34} & 1 & r_{13,14} & r_{24,12} \\ r_{31,24} & r_{32,14} & 1 & r_{34,12} \\ r_{41,23} & r_{42,13} & r_{34,12} & 1 \end{bmatrix}$$

donde, por ejemplo, $r_{12,34}$ es la correlación entre las variables 1 y 2 cuando eliminamos el efecto de la 3 y la 4, es decir, cuando las variables 3 y 4 permanecen constantes.

Puede demostrarse que el coeficiente de correlación parcial entre dos variables es proporcional al coeficiente de una regresión entre las dos variables que incluye también al resto de las variables. En concreto, por ejemplo:

$$r_{12,34} = \widehat{\beta}_{12,34} \left(\sqrt{\widehat{\beta}^2_{12,34} + s^2_{12,34} \left(n - k - 1 \right)} \right)$$

donde k es aquí igual a 4 y $\widehat{\beta}_{12,34}$ se obtiene a partir de la recta de regresión

$$\widehat{x}_1 = \widehat{\beta}_0 + \widehat{\beta}_{12,34} x_2 + \widehat{\beta}_{13,34} x_3 + \widehat{\beta}_{14,34} x_4$$

siendo $s^2_{12,34}$ es la varianza estimada del coeficiente $\widehat{\beta}_{12,34}$ en esta ecuación.

En SPSS se calcula con los menús:

`Analizar -> Correlaciones -> Parciales`

En R se calcula cargando antes la librería `corpcor` y usando el comando `cor2pcor` sobre una matriz de correlaciones habitual, o bien el comando `pcor.shrink` directamente sobre los datos.

Ejemplo

La matriz de correlación para las 7 variables físicas del ejemplo previo, manteniendo el orden de las variables es

$$R = \begin{bmatrix} 1 & 0,83 & 0,93 & 0,91 & 0,84 & 0,59 & 0,84 \\ 0,83 & 1 & 0,85 & 0,82 & 0,84 & 0,62 & 0,72 \\ 0,93 & 0,85 & 1 & 0,85 & 0,80 & 0,55 & 0,85 \\ 0,91 & 0,82 & 0,85 & 1 & 0,80 & 0,48 & 0,76 \\ 0,84 & 0,84 & 0,80 & 0,80 & 1 & 0,63 & 0,63 \\ 0,59 & 0,62 & 0,55 & 0,48 & 0,63 & 1 & 0,56 \\ 0,84 & 0,72 & 0,85 & 0,76 & 0,63 & 0,56 & 1 \end{bmatrix}$$

Se observa que la máxima correlación aparece entre la primera y la tercera variable (estatura y longitud del pie) siendo 0,93. La mínima correlación es entre la longitud del brazo y el diámetro del cráneo (0,48). En general, las correlaciones más bajas aparecen entre el diámetro del cráneo y el resto de las variables.

La Varianza Generalizada

Una medida global escalar de la variabilidad conjunta de k variables es la *varianza generalizada*, que es el determinante de la matriz de varianzas y covarianzas. Su raíz cuadrada se denomina *desviación típica generalizada*, y tiene las propiedades siguientes:

(i) Está bien definida, ya que el determinante de la matriz de varianzas y covarianzas es siempre mayor o igual que 0.

(ii) Es una medida del área (para $k = 2$), volumen (para $k = 3$) o hipervolumen (para $k > 3$) ocupado por el conjunto de datos.

Por ejemplo, supongamos el caso $k = 2$; así, $\mathbf{S}$ puede escribirse como:

$$\mathbf{S} = \begin{bmatrix} s_x^2 & r s_x s_y \\ r s_x s_y & s_y^2 \end{bmatrix}$$

y la desviación típica generalizada es:

$$|\mathbf{S}|^{1/2} = s_x s_y \sqrt{1 - r^2}$$

Si las variables son independientes, la mayoría de sus valores estarán dentro de un rectángulo de lados $6s_x$, $6s_y$ ya que, por el teorema de Tchebychev, entre la media y 3 veces la desviación típica debe estar aproximadamente al menos el 90 % de los datos. En consecuencia, el área ocupada por ambas variables es directamente proporcional al producto de las desviaciones típicas.

Si las variables están relacionadas linealmente y el coeficiente de correlación es distinto de cero, la mayoría de los puntos tenderán a situarse en una franja alrededor de la recta de regresión y habrá una reducción del área tanto mayor cuanto mayor sea r. En el límite, si $r = 1$, todos los puntos están en una línea, hay una relación lineal exacta entre las variables y el área ocupada es cero. La última fórmula describe esta contracción del área ocupada por los puntos al aumentar el coeficiente de correlación.

Análogamente, en el caso tridimensional,

$$|S|^{1/2} = s_x s_y s_z (1 + r_{12}^2 (r_{13} - 1) + r_{13}^2 (r_{12} - 1) - r_{13} r_{12})^{1/2}$$

si las variables no están correlacionadas, el volumen ocupado es proporcional al producto de las desviaciones típicas. Esta cantidad se reduce ante la presencia de correlación como se muestra en la fórmula anterior.

En resumen, análogamente a cómo la desviación típica describe la dispersión de una variable, la desviación típica generalizada describe la dispersión conjunta de un grupo de variables, que depende de la correlación entre ellas.

Ejemplo

Partiendo de la matriz de covarianza **S** de la tabla de datos anterior se tiene que la varianza generalizada viene dada por:

$$|S|^{1/2} = 0{,}0195$$

Como la varianza generalizada mide el grado de dispersión en el espacio, notamos que esta no es muy alta, por otro lado, las correlaciones entre las variables tampoco son muy altas.

Representaciones Gráficas

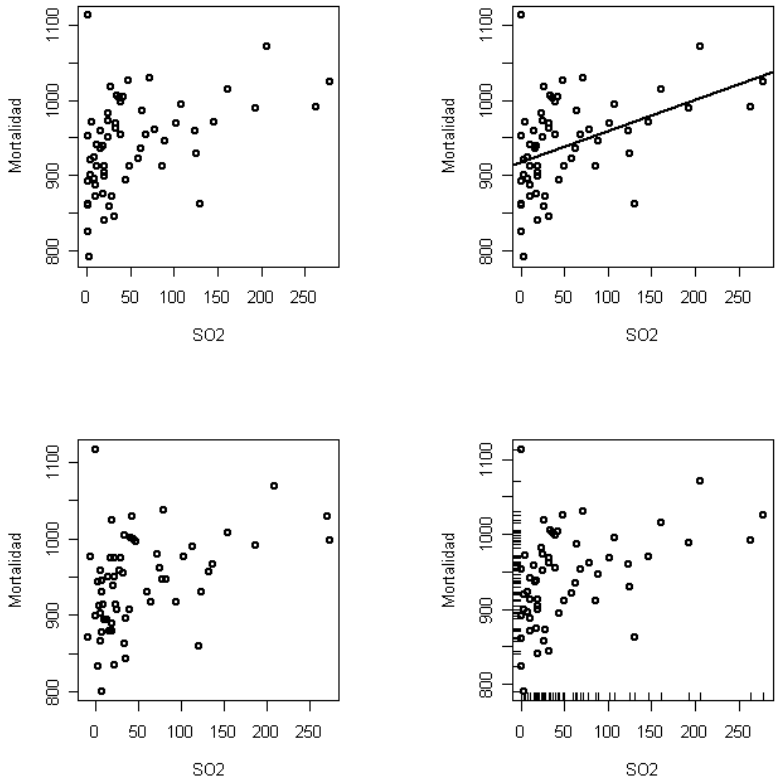
Además de las representaciones univariantes tradicionales, es conveniente representar los datos multivariantes conjuntamente. Para variables discretas podemos construir diagramas de barras tridimensionales, pero no es posible extender la análoga a más dimensiones. Igualmente, podemos construir los equivalentes multidimensionales de los histogramas, pero estas representaciones no son útiles para dimensiones superiores a tres.

Por ejemplo, supongamos unos datos recogidos sobre la cantidad de polución por dióxido de sulfuro y la mortalidad

(ver `http://biostatistics.iop.kcl.ac.uk/publications/everitt/`)

	Lluvia	Educacion	Popden	Noblancos	NOX	SO2	Mortalidad
akronOH	36	11.4	3243	8.8	15	59	921.9
albanyNY	35	11	4281	3.5	10	39	997.9
allenPA	44	9.8	4260	0.8	6	33	962.4
...	...	...	...	...	...	...	...
worctrMA	45	11.1	3678	1	3	8	895.7
yorkPA	42	9	9699	4.8	8	49	911.8
youngsOH	38	10.7	3451	11.7	13	39	954.4

Se pueden considerar las siguientes variaciones sobre gráficos bidimensionales clásicos:



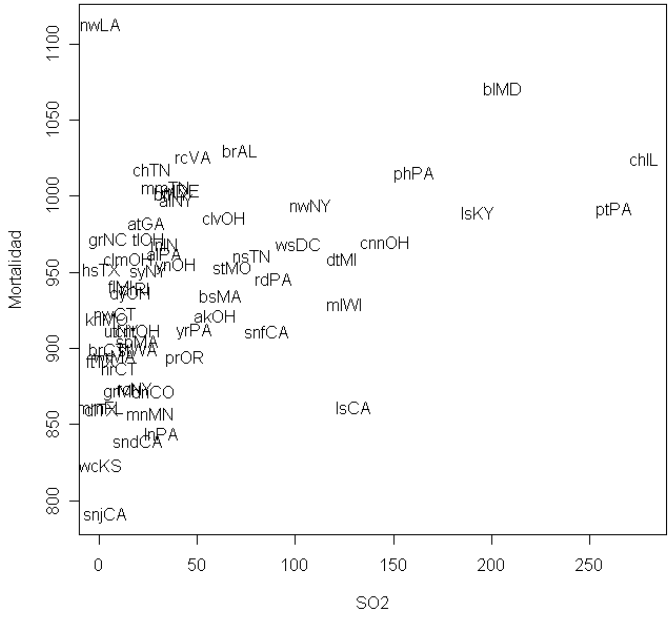
En (a) se presenta el diagrama de dispersión de *mortalidad* frente a *SO2.* En (b) se presenta el mismo gráfico junto con una recta de regresión añadida. En (c) se presenta el mismo diagrama de dsipersión con ruido añadido. En (d) se dibuja también la distribución marginal de cada variable.

El código en R es:

```
par(mfrow=c(2,2))
par(pty="s")
plot(SO2,Mortalidad,pch=1,lwd=2)
title("(a)",lwd=2)
plot(SO2,Mortalidad,pch=1,lwd=2)
abline(lm(Mortalidad SO2),lwd=2)
title("(b)",lwd=2)
```

```
airpoll1<-jitter(cbind(SO2,Mortalidad,50))
plot(airpoll1[,1],airpoll1[,2],xlab="SO2",ylab="Mortalidad",pch=1,lwd=2)
title("(c)",lwd=2)
plot(SO2,Mortalidad,pch=1,lwd=2)
rug(jitter(SO2),side=1)
rug(jitter(Mortalidad),side=2)
title("(d)",lwd=2)
```

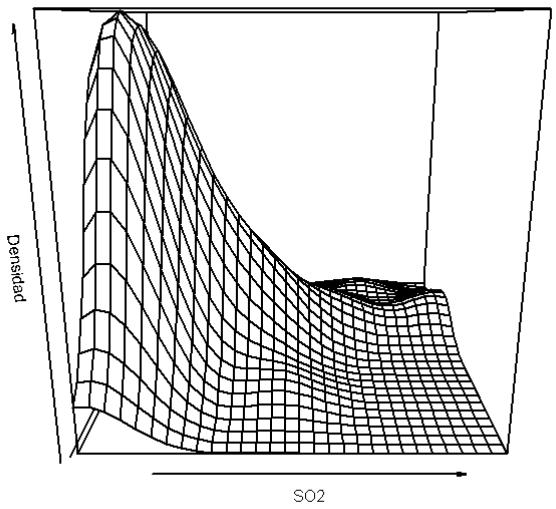
Se puede considerar también un gráfico de dispersión con los nombres de cada una de las observaciones:



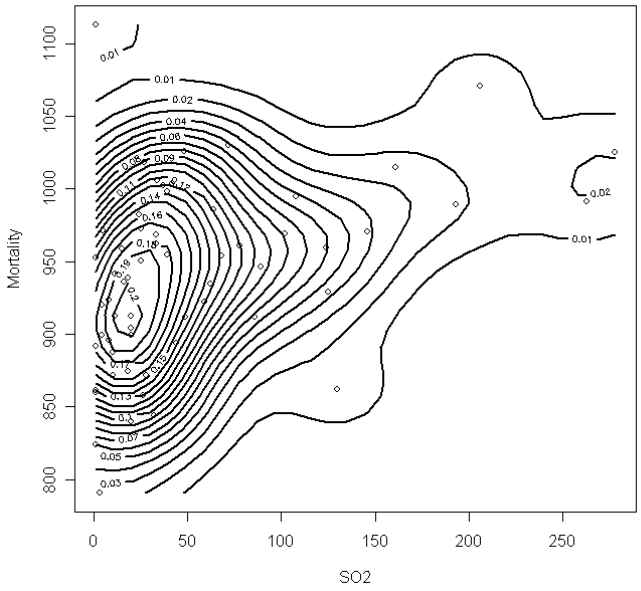
que se construye con el siguiente código:

```
names<-abbreviate(row.names(airpoll))
plot(SO2,Mortalidad,lwd=2,type="n")
text(SO2,Mortalidad,labels=names,lwd=2)
```

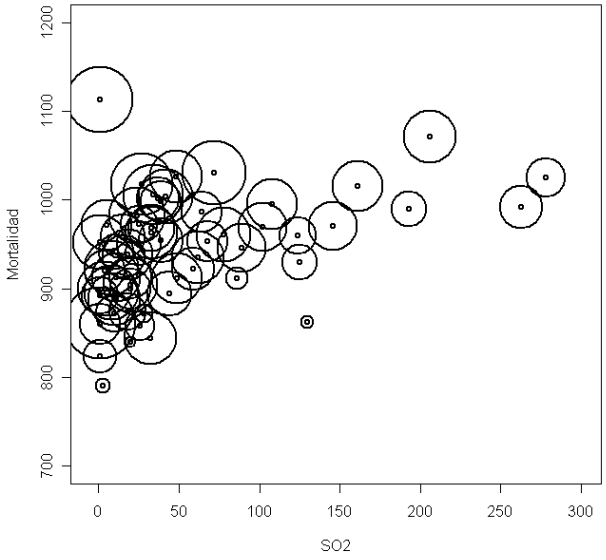
Se pueden considerar histogramas bidimensionales y gráficas de densidad:



y un gráfico de contorno:



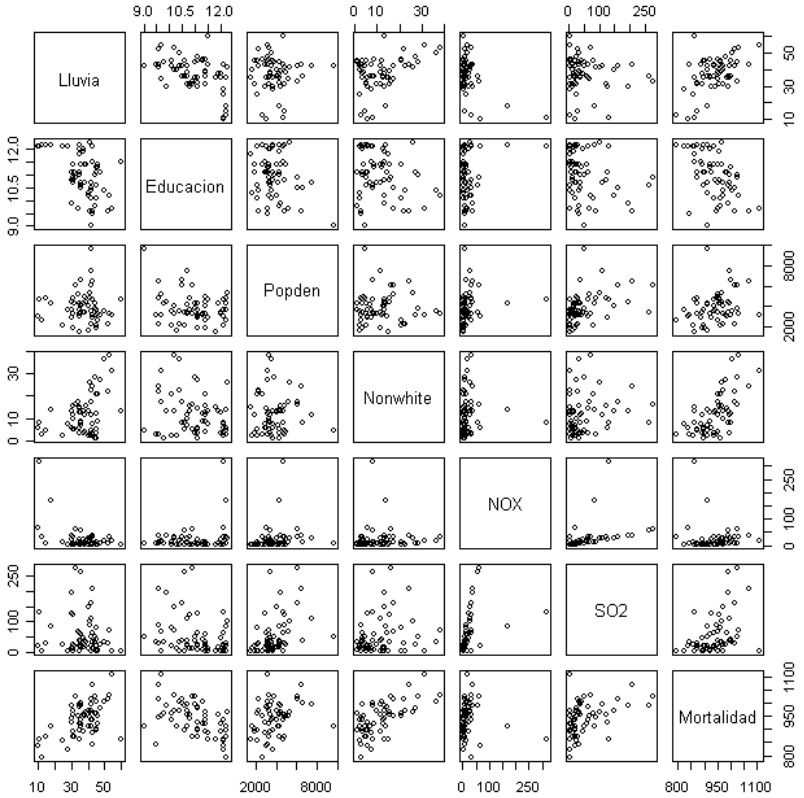
Se puede representar más de una variable mediante gráficos de burbujas:



cuyo código en R es:

```
plot(SO2,Mortalidad,pch=1,lwd=2,ylim=c(700,1200),xlim=c(-5,300))
symbols(SO2,Mortalidad,circles=Lluvia,inches=0.4,add=TRUE,lwd=2)
```

Un gráfico multivariante muy extendido es el de la matriz de dispersión, en el que se cruzan todas las variables entre sí:

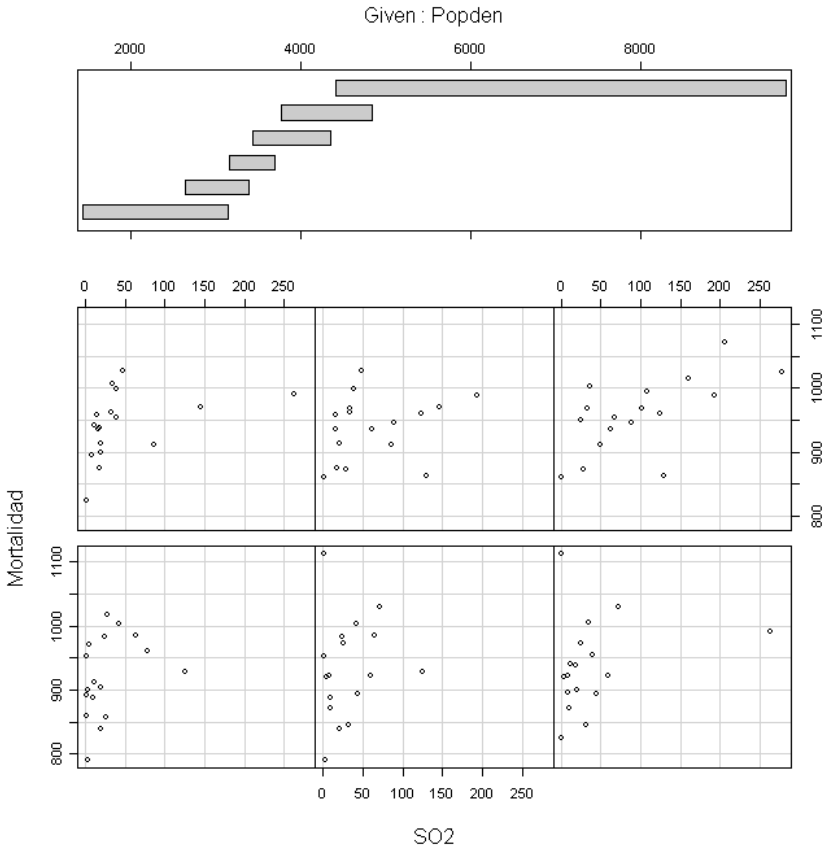


cuyo código en R es simplemente:

```
pairs(airpoll)
```

El gráfico *condicionado* es una herramienta muy útil para visualizar las relaciones entre las variables, condicionadas al valor de otras variables. Se pueden observar, así, relaciones y dependencias entre las mismas.

Por ejemplo el gráfico de mortalidad frente a SO2 condicionado a los valores de densidad de población, es:

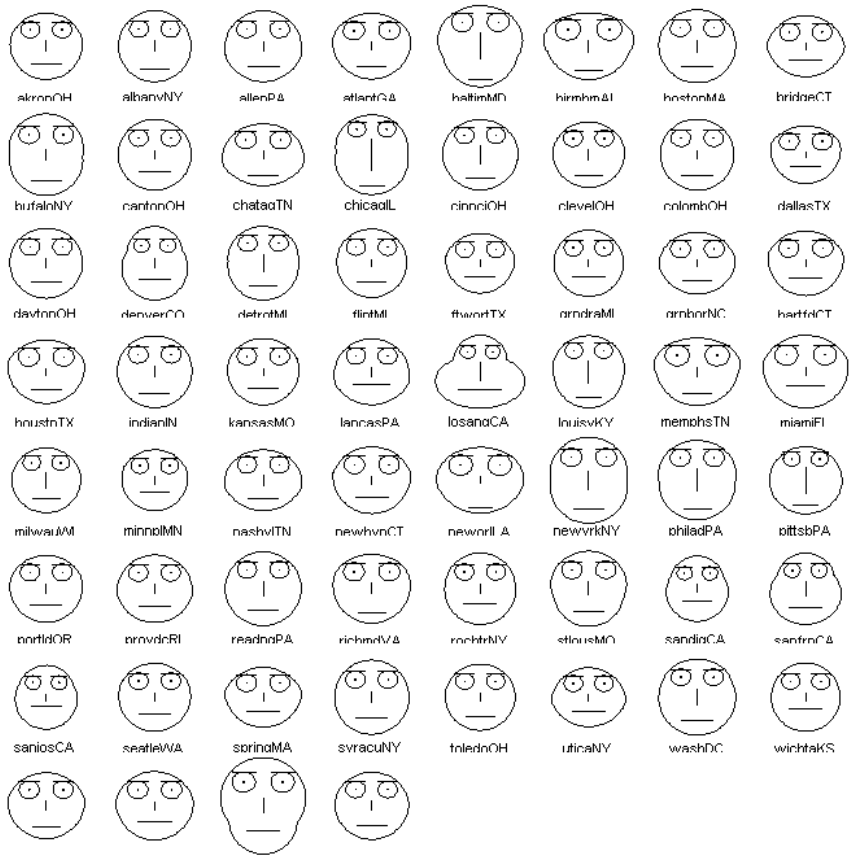


cuyo código en R es simplemente:

```
coplot(Mortalidad~SO2|Popden)
```

Las 6 gráficas en las que se divide la gráfica principal, se deben observar desde abajo y de izquierda a derecha. Cada una de las 6 subgráficas indica la relación que existe entre las variables *Mortalidad* y *SO2* cuando la variable *Popden* tiene los valores que se indican en las barras horizontales del panel de la parte superior.

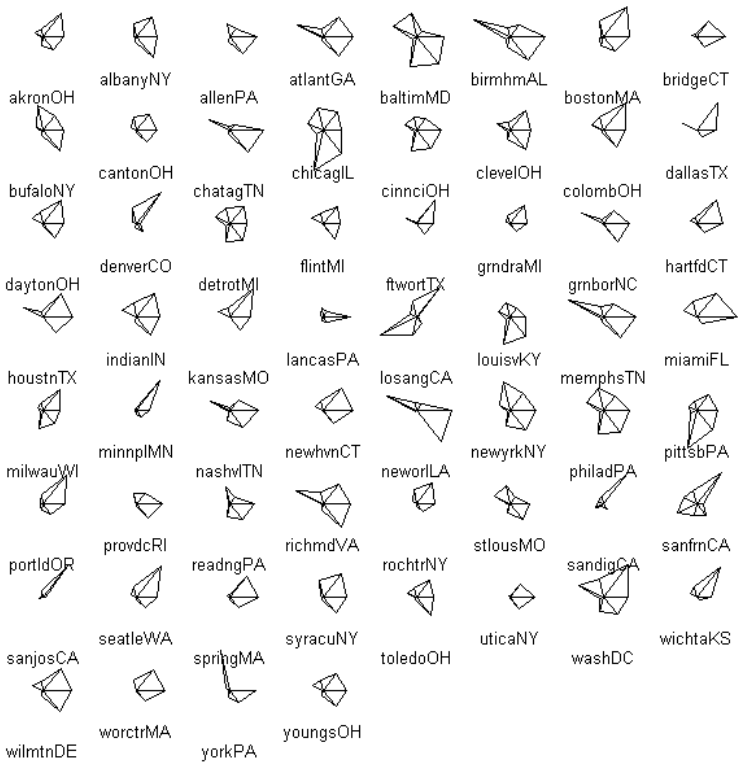
Finalmente, hay gráficas muy populares como las caras de Chernoff y las gráficas de estrellas, donde se asocia a cada variable o bien un rasgo de una cara (en vista de la facilidad con que distinguimos facciones) o bien parte de una estrella:



cuyo código es:

```
library(TeachingDemos)
faces2(airpoll)
```

El gráfico de estrellas, asociado a las observaciones recogidas es:



cuyo código es:

```
stars(airpoll)
```