2. La matriz de datos en ecología: análisis y tratamiento de la información

Ecología Metodológica y Cuantitativa (5C1)
Departamento de Ecología e Hidrología

Curso 2008–09

Índice

1. Introducción

Durante esta sesión se abordarán los aspectos relacionados con la organización y tratamiento de la información procedente de la investigación. Se evaluarán los distintos elementos y procedimientos para construir y preparar una tabla de datos a fin de que, posteriormente, sea posible realizar el adecuado análisis de los datos obtenidos en la investigación.

2. La matriz de datos

La información procedente de la investigación se organiza en forma de tabla o matriz de datos. Los elementos de esta matriz de datos son los valores asociados a: 1) observaciones o individuos, que se corresponden con las filas (siendo el total n); y 2) variables que se corresponden con las columnas (siendo el total p). Cuando el número de variables es uno hablamos de una observación *univariante*; cuando es de dos: *bivariante*; tres: *trivariante*; y en general para más de tres hablamos de observaciones *multivariantes*.

2.1. Las filas de la matriz

En la literatura científica las observaciones pueden recibir distintas denominaciones: objetos, individuos, muestras, quadrats, localidades, etc. La más inconveniente es la tercera de ellas, dado en estos casos el término "muestra" —que, suele indicar una porción extraída: suelo, tejido, agua, sedimento, etc.— puede confundirse con el "todo": el conjunto de individuos u observaciones consideradas en el estudio.

Siguiendo con el ejemplo de la sesión práctica anterior, en una primera aproximación al problema, el investigador ha considerado cada objeto de estudio como un organismo, o ejemplar de la especie de estudio. A tenor de las hipótesis planteadas y con el fin de obtener más información de campo decide considerar como

unidad de estudio una porción del área de estudio: una parcela de 10x10 metros. Inicialmente, divide el área de estudio en 100 parcelas que recubren completamente el área de estudio, tal como muestra la figura 1. Las 100 unidades de muestreo se pueden "etiquetar" o nombran en relación con su posición en latitud y longitud (ver figura 1).

2.2. Las columnas

Una variable es la expresión de una característica observable del objeto. Se consideran distintos tipos de variables atendiendo al conjunto de valores que puedan tomar (conjunto que en estadística se denomina espacio muestral):

- **cualitativas:** Cuando entre los posibles valores no se puede establecer más que relaciones de igualdad u orden. Cada uno de los posibles valores se denomina nivel o modalidad. Se pueden considerar dos tipos:
  - **nominales:** desde el caso más sencillo en que presentan sólo dos modalidades (también llamado de presencia-ausencia), hasta variables con un gran número de modalidades (p. ej.: tipos de suelo, tipos de comunidad vegetal, ...). En este último caso pueden convertirse en un conjunto de variables de presencia–ausencia o variables *dummy*
  - **ordinales:** las distintas modalidades pueden ser ordenadas por un criterio de "mayor que" o "más que" (p. ej.: ausente, raro, escaso, frecuente, abundante, muy abundante).
- **cuantitativas:** Los valores del espacio muestral permiten comparar valores y tienen sentido las sumas y los productos, así como los cocientes y los porcentajes.
  - **discretas** : Los valores de la variable tan sólo pueden ser enteros; a menudo reflejan conteos: número de individuos por unidad de muestreo, número de descendientes por pareja, etc. Si el número de valores posibles es muy grande cabe una aproximación razonable a una situación continua.
  - **continuas** : Las variables pueden tomar cualquier valor de un intervalo dado (pesos, cantidad de biomasa por unidad de superficie, etc.).

También, puede utilizarse otra forma de agrupar las variables considerando otros criterios: *variable de escala nominal, ordinal, de intervalo y de razón*. Las dos primeras coinciden con la clasificación anterior. Las escalas de intervalo son aquellas en las que la unidad de medida es constante; esto es, la misma diferencia existe entre 3 y 4 que entre 6 y 7. Por ejemplo, temperaturas, altitudes, etc.; en general son aquellas donde el valor cero de la escala es arbitrario. En las escala de razón, a diferencia de lo anterior, el cero tiene significado y pueden expresarse los valores como porcentajes: porcentaje de incremento de biomasa, de número de individuos, ...

Con el fin de discutir algunos aspectos relacionados con el procesos de la medida, y desde un punto de vista puramente estadístico, podemos definir *experimento* como: *el procedimiento o mecanismo por el que asignamos a una observación un elemento de un conjunto de símbolos preestablecido*. Ese conjunto determina el tipo de variable: para las cualitativas una relación de dos o más modalidades; para las cuantitativas el conjunto o un subconjunto de los números enteros (discretas) o reales (continuas). El procedimiento de asignar el valor a la observación está afectado por exactitud y precisión.

Hablamos de exactitud cuando el valor asignado coincide con el verdadero valor. Para realizar medidas exactas los aparatos han de estar bien calibrados. En caso de una mala calibración se produce un error sistemático, que en ocasiones puede corregirse *a posteriori*. Se habla entonces de un sesgo en la medida que no debe confundirse con un sesgo en la estimación (concepto ligado a la inferencia estadística).

Hablamos de precisión cuando distintas medidas dan como resultado prácticamente el mismo valor. Cuanto mayores son las diferencias mayor es el error en la medida. Este concepto está relacionado con el error estadístico en la estimación de un parámetro poblacional, pero no ha de confundirse con él. Desde un punto de vista práctico la precisión puede quedar afectada por una sensibilidad desigual del instrumento de medida en el posible rango de la variable.

La calidad de la medida, precisa y exacta, aumenta el coste de la investigación. No obstante unos datos de "calidad" permiten utilizar pruebas estadísticas "fuertes", y obtener así resultados concluyentes.

Ejercicios 1:

Siguiendo con el ejemplo, y considerando las hipótesis planteadas anteriormente, el investigador decide obtener las siguiente información para cada parcela:

- La abundancia de la especie de matorral en estudio.
- La abundancia de individuos de la especie muertos.
- La abundancia de la especie competidora.
- La abundancia del insecto depredador.
- La localización de la parcela: latitud, longitud y cota.
- La profundidad del nivel freático.

A tenor de esta información:

1. Describir los 5 componentes de cada una de las variables que ha de medirse (nombre, símbolo, procedimiento de muestreo, valores, unidades). ¿Cuantas unidades de muestreo hay?
2. ¿Puede, desde el punto de vista de la naturaleza o propiedades descritas, hablarse de dos grandes tipos de variables? ¿Cuáles?

2.3. La notación

En adelante, usaremos la siguiente notación: definida la matriz de datos como una matriz de n individuos por p variables, y hablaremos, en general, del valor x_{ij} para el individuo i–ésimo y la variable j–ésima:

	v_1	v_2	$\ldots$	v_j	$\ldots$	v_p
w_1	$x_{1,1}$	$x_{1,2}$	$\ldots$	$x_{1,j}$	$\ldots$	$x_{1,p}$
w_2	$x_{2,1}$	$x_{2,2}$	$\ldots$	$x_{2,j}$	$\ldots$	$x_{2,p}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
w_i	$x_{i,1}$	$x_{i,2}$	$\ldots$	$x_{i,j}$	$\ldots$	$x_{i,p}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
w_n	$x_{n,1}$	$x_{n,2}$	$\ldots$	$x_{n,j}$	$\ldots$	$x_{n,p}$

Nota: en algunos casos, por cuestiones ajenas a la propia estructura de los datos, en lugar de representar la matriz de datos se representa su traspuesta (p. ej.: para ahorrar espacio en una presentación de resultados).

3. Las matrices en R

El caso de una matriz es similar al de un vector, sólo que ahora disponemos de dos subíndices: uno para indicar el subíndice de la fila y otro para el de la columna, puede construirse una matriz a partir de un fichero (mediante la función `read.table()`), o de los datos de un vector usando la función `matrix(v,f,c)`, donde, `v` es el vector, `f` es el número de filas de la matriz `c` es el número de columnas. Si, para simplificar, sustituimos un vector por una serie, podemos construir una matriz `m`:

```
matrix(1:12,4,3) ->m
```

Para referirnos a la matriz o partes de ellas utilizaremos una notación semejante a la empleada con vectores, teniendo en cuenta las dos dimensiones implicadas en la matriz. Así:

`m` y `m[]`	implican toda la matriz.
`m[1,]`	indica todos los elementos de la primer fila.
`m[,3]`	indica todos los elementos de la tercer columna.
`m[1:2,2:3]`	submatriz con los elementos de las dos primeras filas para la segunda y tercera columnas.

Para el manejo aritmético de matrices existe funciones específicas así como operadores especialmente destinados; por ejemplo, el producto de dos matrices puede obtenerse con el operador `%*%`.

Tabla 1. Ejemplo de hoja de campo.

Fecha		Localidad		Muestreador	
Pedregosidad		Pendiente		Prof. freático	
Mat litológico		Cóncavo/convexo		Encharcamiento	
Rosmarinus officinalis			*Thymelaea hirsuta*		
Rhamnus oleoides			*Salsola kali*		

...

...		...	

Ejercicios 2:

1. Construir la matriz `m` con la ayuda de la función `matrix()` a partir de los siguientes datos:

```
 7  9 1
21  6 6
 1 36 5
22 20 3
```

2. ¿Cuál es el resultado de la expresión `m * 2`? ¿por qué?
3. ¿Cuál es el resultado de la expresión `t(m)`? ¿Qué efecto tiene aplicar la función `t()` a una matriz?
4. ¿Qué ocurre al utilizar las siguientes expresiones:?

   `plot(m)` `plot(m[,1:2])` `plot(m[,1:2],col=m[,3])`
5. Para dar nombre a las columnas (las variables de una matriz de datos) puede utilizarse una expresión como la siguiente: `colnames(m)<-c("v1","v2","v3")`. Utilizar la función `rownames()` para etiquetar las filas.

4. Mecanización de la información: ficheros de datos

En la práctica, como resultado de la investigación, disponemos de un conjunto de datos procedentes de la observación directa en el campo: las medidas *in situ*; los resultados analíticos de muestras de suelo o tejidos; del conteo, tras la separación del material procedente de las trampas de capturas; la información de estaciones termopluviométricas, etc.

La información, habitualmente, queda registrada en *hojas de campo* u *hojas de trabajo* pensadas para anotar la información de forma cómoda y organizada (Tabla 1). Llamamos *mecanización* al proceso de traslado de la información a un sistema informático: los datos se organizan en un fichero que se almacena sobre un soporte magnético u otra forma de almacenamiento digital.

Para realizar la mecanización es preciso utilizar algún programa y ciertas reglas de organización de la información. Lo más recomendable es el uso de programas generales (editores, hojas de cálculo, sistemas de bases de datos, etc.) dependiendo de la magnitud del volumen de información.

La opción más sencilla consiste en la utilización de un editor. En él se considerará cada línea como una línea de la matriz de datos y cada columna debe estar representada por un valor —en la terminología informática a esto se denomina registros y campos—. Los valores de las distintas columnas se separan mediante un carácter al que se otorga el papel de *separador de campos* (*field separator*). Se utilizan distintos caracteres como separador de campo: espacio, la coma, el punto y coma, el tabulador, etc. Los datos se almacenan en formato ASCII y pueden ser leídos por distintos programas sin problemas.

Opcionalmente se puede añadir una primera línea, denominada cabecera (*header*) en la que se escriben los nombre de las columnas o variables. Para evitar problemas con caracteres especiales se entrecomillan estos

nombres. Se puede proceder de forma análoga para incluir los nombres de los objetos; para ello se incluyen al principio de la línea.

Para evitar problemas se suelen codificar todas las variables utilizando en estos ficheros únicamente valores numéricos (salvo en los correspondientes a las etiquetas de filas y columnas).

Debe tenerse en cuenta un caso especial: cuando el investigador carece de un dato, es decir, se ha perdido esta información y no es posible disponer de ella. Para estos casos: valores perdidos (*missing values*) se utiliza una codificación especial, bien con un valor fuera del rango de valores posibles (p. ej.:-999.999,-1, ...), por una cadena de caracteres de significado especial: `M`, `Na`, etc.

La utilización de hojas de cálculo puede ser ventajosa en algunas ocasiones, los datos pueden llevarse a un formato ASCII mediante la exportación como fichero `.csv` (*comma separated values*), es decir, valores separados por comas. En otras ocasiones resulta más conveniente utilizar un proceso optimizado para la edición de los datos; un ejemplo típico es la información de presencia–ausencia, que en una hoja de cálculo se manejan con una cierta incomodidad.

Ejercicios 3:

1. Utilizando un programa para la consulta de documentos web, acceder a la página oficial de la asignatura *Ecología Metodológica y Cuantitativa*

   `http://www.um.es/docencia/emc`

   y consultar y descargar, los fichero de datos disponibles para esta sesión práctica (`din.pob.dat`, `iris.dat` y `evolpobmun.dat`).

2. Describir brevemente considerando la estructura, número de filas y columnas, separador de campo, presencia de línea de cabecera, etc. los ficheros descargados en el punto anterior.

	n	p	cabecera	separador
`din.pob.dat`				
`iris.dat`				
`evolpobmun.dat`				
`territorios.dat`				

4.1. Importación de datos en R

En R, para leer un fichero que contenga una matriz de datos se utiliza la función `read.table()`. Esta función considera, por defecto, como separador de campo el espacio en blanco y no considera la existencia de cabecera con el nombre de variables. Por ejemplo, para leer el fichero `misdatos.dat` que dispone de una línea de cabecera, y se utiliza la coma como separador almacenándolo en la variable `datos`, se utilizaría la expresión: `read.table("misdatos.dat",sep=",",header=T)->datos` (el nombre del fichero puede ser una URL (*uniform resourde locator*), por ejemplo: `http://www.um.es/docencia/emc/datos/din.pob.dat`).

Dado un fichero de datos, si el número de campos de la cabecera es una unidad menor que el de las siguientes líneas (que han de tener obligatoriamente el mismo número de campos), se considera que la primera línea contiene las etiquetas de las variables y el primer campo de la línea se corresponde con la etiqueta de esta.

Ejercicios 4:

1. Utilizar la función `read.table()` para leer las matrices de datos descargadas de Internet en los ejercicios anteriores y asignarlas a una variable que coincida con el nombre (excluida la extensión) del fichero.

2. Utilizar las funciones `ncol()` y `nrow()` para averiguar el número de filas y columnas de las distintas matrices.

3. Utilizar las funciones `colnames()` y `rownames()` para determinar el nombre de las filas y las columnas de las distintas matrices.

4. Utilizar la función `ls()` para averiguar los objetos disponibles. Utilizar la función `attach()` con la matriz `iris`, ¿qué objetos tenemos ahora disponibles? ¿qué sucede si tratamos de usar la variable `lonsep`? ¿qué ocurre si utilizamos `detach(iris);ls()` ¿por qué?

---

5. Tipos de matrices de datos

Tal como hemos discutido hasta ahora, las matrices de datos tienen diferente naturaleza, en relación con los objetivos de la investigación y el procedimiento a utilizar.

Podemos encontrar varios tipos de matrices de datos: donde las columnas son variables propiamente dichas (cuantitativas o cualitativas), matrices donde los elementos representan frecuencias —en estos casos puede considerarse la matriz como una tabla de contingencia, matrices heterogéneas donde se combinan distintos tipos de variables. En otros casos pueden utilizarse matrices de correlación entre columnas de afinidad entre filas.

En ecología suele darse, con independencia de las propiedades de las variables, la existencia de distintos tipos de matrices de datos, principalmente, la que llamamos matriz de datos biológicos (abundancia, presencia, características de las población de las especies o individuos) y la matriz de datos ambientales (que describe la localidades donde se han tomado los datos biológicos).

Ejercicios 5:

1. Considerando un caso sencillo:

   "Para estudiar el efecto de dos sustratos, A y B, en la abundancia de un organismo se han utilizado 10 unidades de muestreo, situadas aleatoriamente, en cada uno de los tipos de sustrato. En el sustrato **A** tenemos el siguiente número de individuos por unidad de muestreo: 1, 4, 1, 1, 0, 2, 2, 1, 0, 0; y en el **B**: 0, 2, 4, 0, 2, 3, 1, 2, 3, 0."

   Los datos pueden esquematizarse de la siguiente forma:

Sustrato A	1	4	1	1	0	2	2	1	0	0
Sustrato B	0	2	4	0	2	3	1	2	3	0

   ¿Cuantas variables se están considerando en el estudio?

   ¿Cuántas filas y columnas tendría la matriz de datos?

   ¿Cómo puede introducirse la información en R?

   Preparar un fichero de datos, con la ayuda del programa `nano`, que permita el uso de la expresión `read.table ("datos.dat",header=T)`

   Para utilizar `nano` se necesitará una terminal adicional. La siguiente orden `nano datos.dat`, que abre un editor similar al block de notas, permitirá escribir los datos. Para terminar se pulsa CTL+X, o tal cómo lo indica la ayuda del propio programa: ^X.

2. En la figura 1, se recogen distintas posibilidades a la hora de tomar muestras de vegetación sobre un área de estudio: muestreos sistemáticos o aleatorios.

   Suponiendo que se elija un muestreo aleatorio: construir la matriz de datos que refleje la abundancia de la especie y la ubicación espacial de las unidades de muestreo.

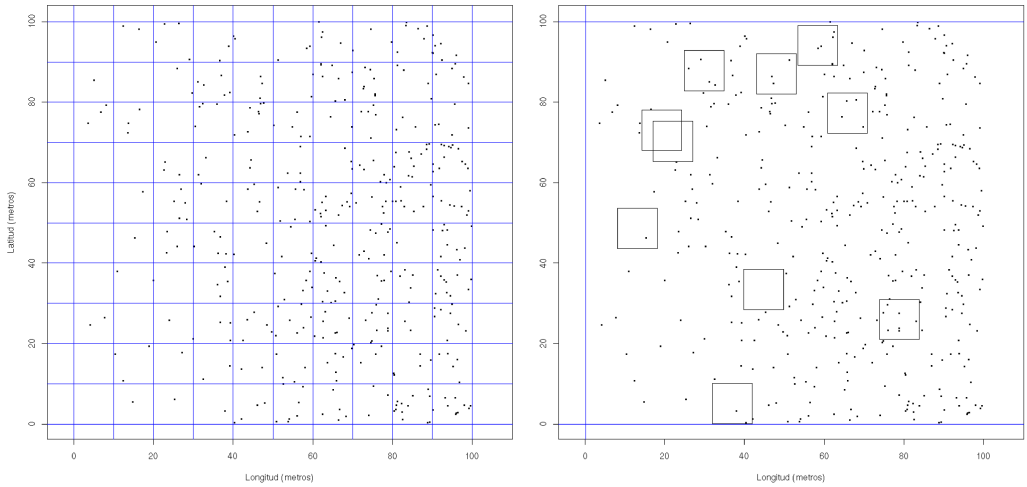

Figura 1: Opciones de muestreo.

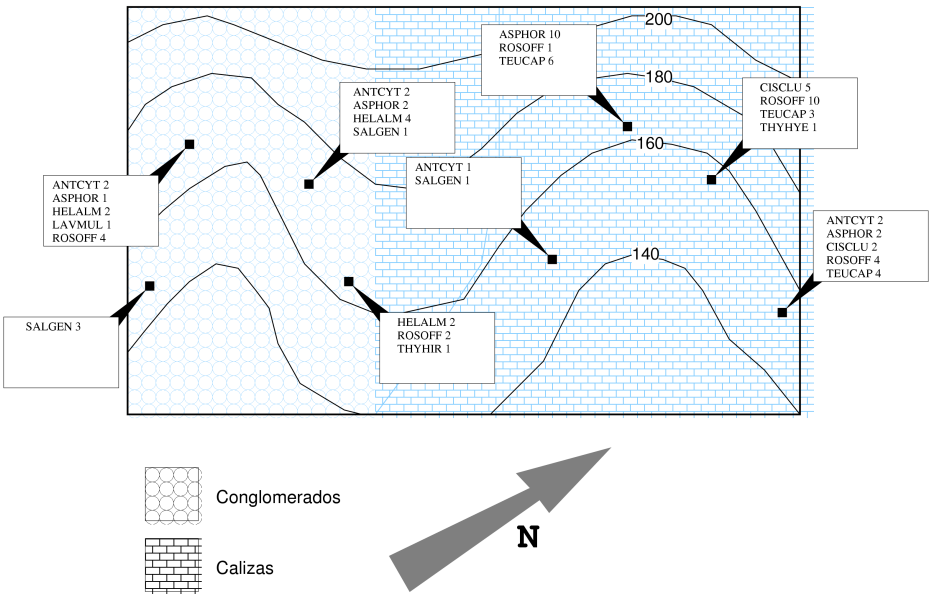

Figura 2: Esquema de muestreo

6. Ejercicios adicionales

1. Localizar 2 trabajos publicados en una revista de Ecología y realizar una descripción de la matriz de datos utilizada por los autores.

2. En el estudio del efecto de un gradiente de orientación en la abundancia de las especies de una comunidad se han tomado 8 unidades de muestreo, tal como refleja la figura 2. Elaborar la cabecera de la matriz de datos de tal manera que pueda construirse ésta quedando recogidas: la orientación de la ladera sobre la que se ubica la unidad de muestreo, el tipo de sustrato y la abundancia de las especies.