Tratamiento de datos con R, STATISTICA y SPSS

CÁSTOR GUISANDE GONZÁLEZ
Catedrático del área de Ecología. Universidad de Vigo (España)

ANTONIO VAAMONDE LISTE
Catedrático del área de Estadística e Investigación Operativa.
Universidad de Vigo (España)

ALDO BARREIRO FELPETO
Investigador del área de Ecología. Universidad de Vigo (España)

Tratamiento de datos con R, STATISTICA y SPSS

DIAZ DE SANTOS

Ediciones Díaz de Santos
E-mail: ediciones@diazdesantos.es
Internet: http://www.diazdesantos.es/ediciones

ISBN: 978-84-7978-998-5
Depósito legal: M. 7.321-2011

Diseño de cubierta: P55 Servicios Culturales
Impresión: FER Fotocomposición
Impreso en España

ÍNDICE

PRÓLOGO Y AGRADECIMIENTOS

Los investigadores necesitan la Estadística para obtener conclusiones válidas a partir de los datos. El método estadístico se ha convertido en parte esencial de la generación de conocimiento científico, y en prácticamente todas las publicaciones especializadas –en casi cualquier campo de conocimiento– las técnicas estadísticas tienen un papel muy importante, hasta el punto de que las revistas científicas suelen disponer de revisores con amplios conocimientos en este campo.

Aunque los usuarios de los métodos de tratamiento y análisis de datos tienen en general conocimientos de Estadística, no suelen ser expertos; para resolver problemas específicos tienen dos opciones: consultar con un especialista –siempre recomendable si el problema es complejo– o recurrir a algún manual de consulta. Frecuentemente los libros de Estadística son excesivamente teóricos, prestando atención únicamente a los aspectos matemáticos, demostraciones, teoremas y propiedades (sin duda de gran importancia para conocer en profundidad las técnicas estadísticas), o bien libros que omiten totalmente los principios básicos y el esquema conceptual que permite comprender el funcionamiento de cada técnica estadística, para centrar todos sus esfuerzos en la aplicación en si misma, presentada como receta al alcance del lector profano.

Los libros más teóricos, pese a su corrección formal, resultan de escasa utilidad para el usuario normal de las técnicas estadísticas, ya que se ve obligado a convertirse en un experto, o al menos a intentarlo, dedicando a ello más tiempo del que dispone, para poder llegar a resolver sus problemas inmediatos de tratamiento de datos, a menudo sin lograrlo. Los libros más aplicados, por el contrario, parecen muy adecuados para el lector que tiene un problema estadístico similar al descrito en el capítulo correspondiente, pero pronto averigua que lamentablemente los problemas son siempre distintos, que existen cuestiones que la receta no trata o no resuelve, y que al utilizarla nunca está a salvo de cometer errores insalvables, que invalidarán los resultados de su investigación.

El principal objetivo de este libro es servir de ayuda a los investigadores, cualquiera que sea su ámbito de conocimiento, que necesitan utilizar técnicas estadísticas para tratar y analizar sus datos. Con un enfoque centrado en la aplicación, se hace énfasis en la explicación de cada método estadístico con un lenguaje accesible, tratando de aclarar de forma breve pero razonada los principios básicos de cada técnica. Numerosas aplicaciones con casos reales son utilizadas para fijar los conceptos y explicar los pasos a seguir en el tratamiento de los datos, desde la introducción de datos hasta la interpretación de los resultados obtenidos.

Se utilizan tres programas estadísticos: SPSS, STATISTICA, y R. Todos ellos son programas de ámbito general muy conocidos y utilizados para el tratamiento estadístico de datos en diversos campos. R es un programa estadístico de uso libre y gratuito, que recientemente ha alcanzado una gran difusión. A la ventaja –nada desdeñable– de su utilización gratuita, une una potencia de análisis superior a la de cualquier otro programa de carácter general, junto con un ritmo de desarrollo de aplicaciones que permite adivinar que pronto desplazará del mercado a los programas más costosos. R es un proyecto abierto y participativo en el

que colaboran –de forma libre y espontánea– miles de investigadores estadísticos, lo que asegura el debate, la crítica, y la actualización permanente de los procedimientos; su único inconveniente, por el momento, es el entorno o interfase de aplicación, con menús de usuario poco (o nada) desarrollados en algunas técnicas, que utiliza un lenguaje de programación que puede parecer complejo a los no iniciados.

Todas las técnicas son aplicadas con los tres programas. Para ello se utilizan datos distintos, aunque a veces el mismo ejemplo es resuelto con dos programas diferentes para que el lector pueda comparar y deducir sus ventajas y limitaciones. Se trata de conseguir que el lector pueda reproducir con sus propios datos cualquier aplicación y, por ello, se sigue siempre un esquema de pasos sucesivos, con el que debe resultar razonablemente sencillo completar un nuevo análisis mediante cualquiera de los tres programas estadísticos mencionados.

Todos los ejemplos se explican detalladamente, para que puedan ser repetidos incluso por el usuario que no tiene conocimientos previos de Estadística ni del programa. Las distintas técnicas estadísticas son presentadas de forma independiente, de manera que el lector interesado en cualquiera de ellas pueda empezar directamente por el capítulo correspondiente.

Los supuestos o hipótesis que cada método requiere son verificados cuidadosamente: es necesario que el lector compruebe también todas las hipótesis en cualquier aplicación de la misma técnica, y lo haga constar así junto a la presentación de los resultados. Los métodos estadísticos solo son válidos cuando se aplican correctamente en su contexto adecuado, y es muy frecuente encontrar aplicaciones desafortunadas en las que obviamente no se cumplen las condiciones necesarias, o en las que simplemente no se han comprobado.

Hemos prestado también especial atención a la interpretación correcta de los resultados obtenidos. Es bastante habitual observar –en la publicación de resultados– cómo se obtienen conclusiones que no pueden deducirse de los datos, generalmente porque se interpreta de forma laxa o generalizada el resultado de una prueba estadística, porque se ignoran las características de los datos utilizados, o porque se desconocen las posibilidades y limitaciones del método estadístico. Aunque en este libro se intenta explicar con claridad cómo deben interpretarse correctamente los distintos resultados, lo cierto es que éste es el aspecto más difícil para el usuario no experto. El único antídoto frente al mal uso de la Estadística consiste en un conocimiento profundo de las técnicas, por lo que recomendamos vivamente que el usuario consulte a un especialista cuando tenga dudas de interpretación en la aplicación de métodos estadísticos complejos.

El libro se acompaña de un CD con los archivos de datos de todos los ejemplos utilizados, así como de los programas de R necesarios para su análisis, con el fin de facilitar su utilización. Una breve guía resumen al final del libro puede utilizarse para una primera orientación sobre la técnica estadística que debe ser utilizada, en función del tipo de datos del que se dispone y de los objetivos buscados con el análisis.

Esta obra no tiene la pretensión de ser exhaustiva. Muchas técnicas simples de Estadística Descriptiva y algunos métodos estadísticos de carácter más específico no son tratados para no aumentar excesivamente su extensión. Hemos tratado de incluir la mayoría de las técnicas de tratamiento estadístico de datos que los investigadores necesitan en su trabajo habitual, añadiendo la resolución de problemas frecuentes que no tienen en otros textos una consideración suficiente, y hemos procurado también situarnos en el lugar del usuario para dar una respuesta satisfactoria a sus demandas, de modo que pueda superar sin dificul-

tades las barreras que a menudo impiden la aplicación de esta herramienta imprescindible que es la Estadística.

Por último, queremos agradecer al profesor Kenneth Roy Cabrera Torres de la Universidad Nacional de Colombia Sede Medellín, su ayuda con los *scripts* de R en los primeros capítulos de este libro.

Capítulo I

REPRESENTACIÓN DE DATOS

I.1. Coordenadas polares

Un paso obligado previo a la realización de cualquier tratamiento de datos es representar los datos gráficamente. Esto es necesario por muchos motivos: por ejemplo, para ver el tipo de relación que existe entre dos variables (lineal, logarítmica, exponencial, etc.); para identificar posibles "*outliers*", datos que son muy diferentes del resto y que se pueden deber simplemente a que hemos introducido mal los datos en el ordenador; para ver el tipo de distribución y variabilidad de los datos, etc. Existen muchos tipos de gráficas que permiten hacer representaciones, que son bien conocidas y de uso común por la mayoría de las personas que trabajan con datos. En este capítulo solo vamos a tratar un tipo de representación que no es tan conocido, el gráfico de coordenadas polares.

Un problema frecuente que surge a la hora de mostrar gráficamente los resultados obtenidos es que es necesario representar más de dos ejes de coordenadas. Sin embargo, en un plano bidimensional lo máximo que se puede dibujar son tres ejes. Las coordenadas polares permiten representar en un gráfico bidimensional cualquier número de ejes de coordenadas.

Supongamos que se quiere representar el nicho de varias especies de árboles considerando como variables del nicho el pH medio del suelo en el que aparece la especie, la temperatura media, la salinidad media del suelo, la pluviosidad media y la concentración media de nitratos en el suelo. Las coordenadas polares consideran las diferentes variables como vectores de fuerza, de tal forma que las coordenadas polares *X* e *Y* de cada especie serían el punto de equilibrio de todos los vectores es decir, de todas las variables (Figura I.1).

El módulo de cada vector sería el valor de la variable y el ángulo de cada vector tendría un valor convenido asignado a cada variable. Por lo tanto, cada especie tendría unas coordenadas *X* e *Y*, lo cual permite su representación en un plano bidimensional y, además, estas coordenadas *X* e *Y* vendrían determinadas por el efecto conjunto de todas las variables consideradas para definir el nicho de las especies. Es decir, cada especie ocuparía una posición en el plano en función de las variables del nicho, lo cual permitiría ver gráficamente que especies tienen un nicho más parecido.

Debido a que la representación polar se basa en vectores de fuerza, solo se puede utilizar con variables que tengan igual peso, es decir, que los valores de las variables puedan ser comparables. En el caso de que las variables no sean equiparables es necesario realizar una estandarización previa de los datos.

En un análisis multivariante de componentes principales, las puntuaciones obtenidas en cada eje para cada caso serían un ejemplo de variables que son en general equiparables. Los porcentajes de alimento consumido por varias especies también serían otro caso de variables equiparables, así como, por ejemplo, el porcentaje granulométrico del suelo.

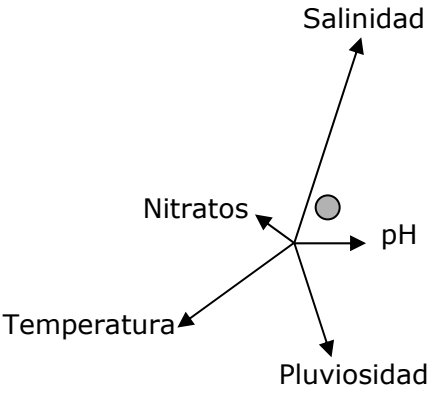

Figura I.1 Esquema de un sistema de coordenadas polares donde el círculo gris indica el centro de gravedad o punto de equilibrio de todas las variables. Las coordenadas de ese punto gris serían las coordenadas *X* e *Y* de la especie.

Por el contrario, si por ejemplo queremos una representación de varias lagunas, para ver gráficamente qué lagunas tienen unas características físico-químicas semejantes, considerando la temperatura, conductividad, pH, oxígeno, etc., estas variables no serían equiparables y sería necesario realizar una estandarización previa.

I.1.1. ESTANDARIZACIÓN DE LOS DATOS

Como ejemplo de estandarización previa al cálculo de las coordenadas polares vamos a trabajar con datos físico-químicos de distintas lagunas, para las que, en algunas de ellas, existen varias estaciones de muestreo.

En la Tabla I.1 (Cuadro I.1) se muestran los datos y en el archivo **Cuadro I.1.xls**, además de los datos, se muestran todos los pasos a seguir para estandarizar los datos, los cuales también se describen en el Cuadro I.1.

En el Cuadro I.2 hay otro ejemplo a modo de práctica, del método para estandarizar datos.

I.1.2. ASIGNACIÓN DE ÁNGULOS A LAS VARIABLES

Un paso previo al cálculo de las coordenadas es asignar ángulos a las variables consideradas. La asignación de ángulos a las variables es diferente si hay valores negativos (por ejemplo una salida de un análisis multivariante) o si, por el contrario, solo hay valores estandarizados en una escala de 0 a 1 y, por tanto, no hay valores negativos.

En el caso de que solo existan valores positivos en una escala de 0 a 1 (los valores están estandarizados), lo que se hace es restar a todos los valores 0,5 antes de comenzar con el proceso de asignar ángulos a las variables. Si el rango de los datos no es de 0 a 1 (los valores no están estandarizados) ya que, por ejemplo, es una salida de un análisis multivariante que tiene valores negativos, no se resta el valor de 0,5 a todos los datos.

CUADRO I.1. Estandarización de datos

EJEMPLO. La estandarización de los datos físico-químicos de diferentes lagunas se muestra en la siguiente tabla. Los datos están en el archivo **Cuadro I.1.xls**.

Tabla I.1. Datos físico-químicos de distintas lagunas.

Laguna	Estación	Temperatura (ºC)	Conductividad (μS cm^{-1})	pH	Nitrato (μM)	Nitrito (μM)	Amonio (μM)	Fosfato (μM)	Silicato (μM)
1	1	29,1	495	6,08	0,00	1,090	1,140	0,320	62,78
2	1	28,1	1537	7,85	0,52	0,280	0,940	0,384	231,67
	2	28,6	1568	8,52	0,00	0,253	0,839	0,428	262,31
3	1	28,2	755	8,52	0,52	0,275	0,937	0,384	231,67
	2	28,5	739	8,24	0,00	0,348	0,339	0,761	172,69
	3	28,6	748	8,60	0,13	0,386	0,419	0,867	189,41
4	1	27,5	905	7,63	0,00	0,893	0,915	1,056	91,699
5	1	26,4	419	7,72	0,00	0,708	0,668	0,258	113,18
6	1	23,9	1034	7,64	0,42	0,301	1,229	0,263	189,04
7	1	32,6	217	7,42	0,30	0,843	2,076	0,618	118,00
8	1	31,3	371	7,32	0,00	0,697	0,635	0,489	99,225
9	1	32,3	1162	7,17	0,04	0,248	0,762	1,811	326,63

Paso 1. Cálculo de los valores máximos y mínimos de las variables.

	Temperatura (ºC)	Conductividad (μS cm^{-1})	pH	Nitrato (μM)	Nitrito (μM)	Amonio (μM)	Fosfato (μM)	Silicato (μM)
Máximo	32,6	1568	8,6	0,52	1,09	2,076	1,811	326,63
Mínimo	23,9	217	6,08	0	0,248	0,339	0,258	62,78

Paso 2. Estandarización a una escala de 0 a 1 de todas las variables.

A cada uno de los valores de las variables se aplica la siguiente fórmula:

$$VE = \frac{x - Min}{Max - Min}$$

donde *VE* es el valor estandarizado, *Max* y *Min* son los valores máximo y mínimo de cada variable, respectivamente, que se calcularon en el paso 1, y *x* es cada uno de los valores de cada variable. Los valores estandarizados se muestran en la Tabla I.2.

Tabla I.2. Valores estandarizados de los datos físico-químicos que se muestran en la Tabla I.1.

Laguna	Estación	Temperatura (ºC)	Conductividad (μS cm^{-1})	pH	Nitrato (μM)	Nitrito (μM)	Amonio (μM)	Fosfato (μM)	Silicato (μM)
1	1	0,598	0,206	0,000	0,000	1,000	0,461	0,040	0,000
2	1	0,483	0,977	0,702	1,000	0,038	0,346	0,081	0,640
	2	0,540	1,000	0,968	0,000	0,006	0,288	0,109	0,756
3	1	0,494	0,398	0,968	1,000	0,032	0,344	0,081	0,640
	2	0,529	0,386	0,857	0,000	0,119	0,000	0,324	0,417
	3	0,540	0,393	1,000	0,250	0,164	0,046	0,392	0,480
4	1	0,414	0,509	0,615	0,000	0,766	0,332	0,514	0,110
5	1	0,287	0,150	0,651	0,000	0,546	0,189	0,000	0,191
6	1	0,000	0,605	0,619	0,808	0,063	0,512	0,003	0,479
7	1	1,000	0,000	0,532	0,577	0,707	1,000	0,232	0,209
8	1	0,851	0,114	0,492	0,000	0,533	0,170	0,149	0,138
9	1	0,966	0,699	0,433	0,077	0,000	0,244	1,000	1,000

CUADRO I.1. *(Continuación)*

Estadística descriptiva con R

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro I.1.R**, el cual es un archivo *script* que contiene las instrucciones para realizar la estadística descriptiva que se muestra en este ejemplo. Este archivo está en el CD que acompaña al libro en la carpeta Capítulo I. Para grabar los resultados, es mejor pasar el archivo **Cuadro I.1.R** a un directorio del disco duro y seleccionarlo como directorio de trabajo.

Paso 1. Para importar bases de datos desde Excel a R se recomienda guardar el archivo de Excel en formato ".csv" y luego leer desde R utilizando la función read.csv o read.csv2 según el tipo de formato ".csv" usado por Excel para guardar los datos. Las instrucciones para realizar este tipo de importación de datos están en el Apéndice II en el apartado AII.4.2. En este ejemplo usaremos el archivo **Cuadro I.1.V.csv**, en el cual se encuentran los datos mostrados anteriormente en la Tabla I.1.

Paso 2. La siguiente ventana muestra las instrucciones del archivo *script* **Cuadro I.1.R**. Lo único que hay que hacer es cambiar la sección modificable por el usuario, para poder hacer la estandarización de cualquier tipo de archivo de datos.

```
# Cuadro I.1
# Estandarización de datos

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro I.1.V.csv",encoding="latin1")

# Selección de las variables de interés
varInteres<-c("Temperatura","Conductividad","pH","Nitrato",
              "Nitrito","Amonio","Fosfato","Silicato")

# Selección de las variables de identificación

varID<-c("Laguna","Estación")

# Nombre del archivo de salida con los resultados de la
# estandarización
nomSalida<-"Salida Cuadro I.1.V.csv"
```

Paso 3. En primer lugar hay que importar los datos y, por tanto, especificar el nombre del archivo de entrada, en este caso **Cuadro I.1.V.csv**.

datos<-read.csv2("Cuadro I.1.V.csv",encoding="latin1")

CUADRO I.1. *(Continuación)*

Paso 4. Una vez importados los datos hay que definir las variables que se quieren estandarizar.

varInteres<-c("Temperatura","Conductividad",”pH”,"Nitrato", "Nitrito","Amonio","Fosfato","Silicato")

Paso 5. A continuación se seleccionan las variables que se utilizan como códigos de identificación de los casos.

varID<-c("Laguna","Estación")

Paso 6. El último paso que tiene que hacer el usuario es definir el nombre del archivo donde se grabarán los valores estandarizados.

nomSalida<-"Salida Cuadro I.1.V.csv"

Paso 7. Para ejecutar el archivo *script e*s posible posicionarse en cada línea del archivo *script* y pulsar *ctrl-R*, o seleccionar varias líneas o todo el archivo con *ctrl-A* y luego pulsar *ctrl-R*, para ejecutar varias líneas o todo el archivo a la vez. Los resultados se muestran en la siguiente ventana.

```
> resultado<-rbind(maximos,minimos)
>
> # Estandarización a una escala de 0 a 1 de todas las variables dada
> resultado2<-cbind(baseID,t((t(baseVar)-minimos)/(maximos-minimos)))
> if(length(varID)==1) colnames(resultado2)[1]<-varID
>
> ########################################################
> # Sección que muestra los resultados
> ########################################################
>
> # Muestra los resultados de mínimos y máximos
> resultado
        Temperatura Conductividad   pH Nitrato Nitrito Amonio Fosfato Silicato
maximos        32.6          1568 8.60    0.52    1.09   2.08    1.81   326.63
minimos        23.9           217 6.08    0.00    0.25   0.34    0.26    62.78
>
> # Muestra los resultados de la estandarización en la escala 0 a 1.
> resultado2
   Laguna Estación Temperatura Conductividad        pH    Nitrato    Nitrito
1       1        1   0.5977011     0.2057735 0.0000000 0.00000000 1.00000000
2       2        1   0.4827586     0.9770540 0.7023810 1.00000000 0.03571429
3       2        2   0.5402299     1.0000000 0.9682540 0.00000000 0.00000000
4       3        1   0.4942529     0.3982235 0.9682540 1.00000000 0.03571429
5       3        2   0.5287356     0.3863805 0.8571429 0.00000000 0.11904762
6       3        3   0.5402299     0.3930422 1.0000000 0.25000000 0.16666667
7       4        1   0.4137931     0.5092524 0.6150794 0.00000000 0.76190476
8       5        1   0.2873563     0.1495189 0.6507937 0.00000000 0.54761905
9       6        1   0.0000000     0.6047372 0.6190476 0.80769231 0.05952381
10      7        1   1.0000000     0.0000000 0.5317460 0.57692308 0.70238095
11      8        1   0.8505747     0.1139896 0.4920635 0.00000000 0.53571429
12      9        1   0.9655172     0.6994819 0.4325397 0.07692308 0.00000000
       Amonio    Fosfato  Silicato
1  0.45977011 0.03870968 0.0000000
2  0.34482759 0.07741935 0.6400985
3  0.28735632 0.10967742 0.7562251
4  0.34482759 0.07741935 0.6400985
5  0.00000000 0.32258065 0.4165624
6  0.04597701 0.39354839 0.4799318
7  0.33333333 0.51612903 0.1096077
8  0.18965517 0.00000000 0.1910176
9  0.51149425 0.00000000 0.4785295
10 1.00000000 0.23225806 0.2092856
11 0.17241379 0.14838710 0.1381467
12 0.24137931 1.00000000 1.0000000
```