Introducción a R

Email: [delval@decsai.ugr.es](mailto:delval@decsai.ugr.es)

1. Exploratory Data Analysis
   1. *Ejemplo 1, hip dataset*

\* Descargate el dataset hip con el siguiente commando

hip <-read.table("http://astrostatistics.psu.edu/datasets/HIP\_star.dat", header=T,fill=T)

\* Una vez descargado comprueba la dimensión y los nombres de las columnas del dataset. ¿Qué dimensión tiene? ¿qué datos alberga?

\* Muestra por pantalla la columna de la variable RA

\* Calcula las tendencias centrales de todos los datos del dataset (mean, media) utilizando la function apply

\* Haz lo mismo para las medidas de dispersión mínimo y máximo. ¿Seria posible hacerlo con un único comando?¿Que hace la función range()

\* Sin embargo las medidas mas populares de dispersión son la varianza (var()), su desviación standard (sd()) y la desviación absoluta de la mediana o MAD. Calcula estas medidas para los valores de RA

\* Imagina que quieres calcular dos de estos valores de una sola vez. ¿Te serviría este código?

f = function(x) c(median(x), mad(x))

f(hip[,1])

\* ¿Cuál sería el resultado de aplicar apply(hip,2,f)?

\* Vamos a medir la dispersión de la muestra utilizando el concepto de cuartiles. El percentil 90 es aquel dato que excede en un 10% a todos los demás datos. El cuartil (quantile) es el mismo concento, solo que habla de proporciones en vez de porcentajes. De forma que el percentil 90 es lo mismo que el cuartil 0.90. La mediana “median” de un dataset es el valor más central, en otras palabras exactamente la mitad del dataset excede la media. Calcula el cuartil .10 y .50 para la columna RA del dataset hip. Sugerencia: quantile()

\* Los cuantiles 0.25 y 0.75 se conocen como el first quartile y el third quartile, respectivamente. Calcula los cuatro cuartiles para RA con un único comando.

\* Otra medida de dispersion es la diferencia entre el primer y el tercer cuartil conocida como rango intercuartil (IQR) Inter Quantile Range. ¿Obtienes ese valor con la función summary()?

\* Hasta ahora has ignorado la presencia de valores perdidos NA. La función any() devuelve TRUE si se encuentra al menos un TRUE en el vector que damos como argumento. Su combinación con is.na es muy útil. ¿qué obtienes cuando ejecutas el siguiente comando? ¿Cómo lo interpretas?

hasNA = function(x) any(is.na(x))

apply(hip,2,hasNA)

\* Prueba a ejecutar el siguiente comando.

\* Como has observado nos devuelve NA para toda la columna, normalmente querríamos poder usar la función sobre el resto de datos que no son NA: Para ello podemos utilizar la función na.omit. ¿Que ocurre cuando lo hacemos?. Usando apply calcula la media para hip y hip1. Intenta calcular la media de forma que solo cambie la de B.V cuando ignores los valores NA.

\* Obten una idea aproximada de tus datos mediante la creación de un boxplot del hop dataset

\*Crea un scatterplot que te compare los valores de RA y DE. Representa los puntos con el símbolo ‘.’ Y que estos puntos sean de color rojo si DE excede de 0. Sugerencia ifelse()

\* Haz un scatterplot de RA y pmRA. ¿Ves algún patrón?

\* En vez de crear los plots por separado para cada par de columnas, hazlos con un solo comando con el scatterplot matrix

\* Para poder acceder a las variables por su nombre usa attach(hip).Vamos a seleccionar las estrellas Hyadas del dataset aplicando los siguientes filtros:

* RA in the range (50,100)
* DE in the range (0,25)
* pmRA in the range (90,130)
* pmDE in the range *(-60,-10)*
* e\_Plx <5
* Vmag *>4* **OR** B.V *<0.2* (this eliminates 4 red giants)

\* Crea un nuevo dataset con la aplicación de estos filtro. El Nuevo dataset se llama hyades. ¿Que dimensiones tiene? Grafica un scatterplot de Vmag vs B.V

* 1. *Ejemplo 2, iris dataset*

\* Vamos a utilizar el ejemplo del dataset iris que está incluido en la distribución de R. Este dataset fue creado por Douglas Fisher. Consta de tres clases y tipos de 3 clases de tipos de flores:

* \_setosa\_
* \_virginica\_
* \_versicolor\_

Cada una de ellas con cuatro atributos:

* sepal width
* sepal length
* petal width
* petal length

\* Inspecciona las primeras filas del dataset y calcula el summary() del mismo con cada atributo del dataset

\* Crea un histograma de petal.width , teniendo en cuenta que el numero de bins es variable fija este a 9. Añádele color y nombres al eje x "Petal Width"y al gráfico dale el nombre de "Histogram of Petal Width". Crea un histograma para cada variable

\*Crea los cuartiles del dataset

\* Representa en un boxplot la variable de ancho de hoja dependiendo del tipo de hoja que tengan

\* Crea los cuartiles para cada tipo de iris y represéntalos en un plot como líneas cada una de un color

\* Crea los boxplot de la longitud del pétalo en función de la especie de Iris.

\* Compara con scatter plots las variables entre sí.

\* El conjunto de datos “swiss” contiene una medida estandarizada de fecundidad y varios indicadores socioeconómicos para cada una de las 47 provincias francófonas de Suiza.

1. ¿Qué diagrama dibujaría para mostrar la distribución de todos los valores? ¿Qué conclusiones sacarías?

2. Dibuje gráficos para cada variable. ¿Qué puede concluir de las distribuciones con respecto a su forma y posibles valores atípicos?

3. Dibuja un diagrama de dispersión de Fertilidad frente a % Catholic. ¿Qué tipo de áreas tienen las tasas de fertilidad más bajas?

4.¿Qué tipo de relación existe entre las variables Educación y Agricultura?

\* El conjunto de datos de aceites de oliva es bien conocido y se puede encontrar en varios paquetes, por ejemplo, como aceitunas en **extracat.**. La fuente original de los datos es el artículo [Forina et al., 1983].

1. Dibuje un scatterplot de las ocho variables continuas. ¿Cuáles de los ácidos grasos están fuertemente asociados positivamente y cuáles fuertemente asociados negativamente?

2. ¿Hay valores atípicos u otras características que valga la pena mencionar?

\* El conjunto de datos se llama Lanza del paquete HSAUR2.

1. Se informan los datos de cuatro estudios. Dibuje un diagrama para mostrar si los cuatro estudios son igualmente grandes.

2. El resultado se mide por la clasificación de la variable con puntuaciones de 1 (mejor) a 5 (peor). ¿Cómo describirías la distribución?

\* El paquete vcdExtra incluye datos de un viejo estudio de cáncer de mama sobre la supervivencia o muerte de 474 pacientes.

1. Convierta los datos en un data frame y dibuje gráficos para comparar las tasas de supervivencia, primero, por grado de malignidad y, en segundo lugar, por centro de diagnóstico.

2. ¿Qué diagrama dibujaría para comparar las tasas de supervivencia tanto por grado de malignidad como por centro de diagnóstico? ¿Importa el orden de las variables explicativas?

\* Dataset Crabs (del paquete MASS) [Venables y Ripley, 2002]. Los autores inicialmente se transforman a una escala logarítmica y luego escriben que:

“The data are very highly correlated and scatterplot matrices and brush plots [i.e. interactive graphics] are none too revealing.”.

Utilizando gráficos generales, comente si la transformación logaritmica fue una buena idea y si está de acuerdo con su afirmación sobre las correlaciones.

Extra

1. Como crear subgrupos de datos en R

Busca información sobre la function cut(). Para ilustrar su uso vamos a utilizar el dataset state.x77. Si no lo tienes instalado instala el paquete R-Datasets. Usa la función head() para ver como son tus datos..

* Extrae la columna Frost y asigna el resultado a la variable frost
* Tu Nuevo objeto es un vector numérico
* Ahora intenta agrupar los datos en frost en tres niveles. Para crear bins en tus datos puedes utilizar la función cut().
* ¿Que obtienes como nombres de los niveles?
* En la realidad no existen estados que tengan frost en días negativos. Esto es porque R añade un poco de padding. Prueba a solucionar el problema utilizando el parámetro include.lowest=TRUE en cut()
* Los nombres de los niveles no son demasiado informativos, especifica nuevos nombres para los niveles
* Después de este paso has creado un factor que clasifica los estados en bajo, medio y alto según el numero de heladas.
* Ahora cuenta el número de estados que hay en cada uno de los niveles. PISTA: utiliza la función table()