ESTADISTICA DESCRIPTIVA

VICTOR MIGUEL TERRON MACIAS

26/5/2021

CLASE 94. ESTADÍSTICA DESCRIPTIVA

El análisis esztadístico se divide en dos partes:

1. Estadística descriptiva
2. Estaística inferencial

El objetivo del análisis exploratorio de la estadística descriptiva es resumir:

- Análisis exploratorio o descriptivo, si nuestro objetivo es resumir, representar y explicar los datos concretos de los que disponemos. La estadística descriptiva es el conjunto de tecnicas usadas con ese fin.
- Análisis inferencial si nuestro objetivo es deducir(inferir), a partir de estos datos, información significativa sobre el total de la población o las poblaciones de interés. Las técnicas que se usan en este caso forman la estadística inferencial. Intervalos de confianza, intervalos de hipotesis.

En la inferencical el objetivo es deducir, predecir a partir de los datos información significativa, uno no puede hablar de análisis exploratorio sin hablar de análisis inferencial.

ANÁLISIS ESTADÍSTICO DE LOS DATOS

Nos centraremos en entender algunas técnicas básicas de la estadística descriptiva orientadas al análisis de datos.

Estas consistirán en una serie de medidas, gráficos y modelos descriptivos que nos permitirán resumir y explorar un conjunto de datos. **Objetivo final**: Entender los datos lo mejor posible.

TIPOS DE DATOS

Trabajamos con datos multidimensionales: observamos varias características de una serie de individuos.

Se registran en un archivo de ordenador con un formato preestablecido (texto simple, csv, txt).

Una de las maneras de almacenar datos es en forma de tablas de datos. En una tabla de datos de cada columna se expresa una variable, mientras que cada fila corresponde a las observaciones de estas variables para un individuo concreto. * Los datos de una misma columna tienen que ser del mismo tipo por que corresponden a observaciones de una misma propiedad. * Las filas en principio son de naturaleza heterogenea, por que pueden contener datos de diferentes tipos.

TIPOS DE DATOS

Los tipos de datos que consideramos son los siguientes:

- Datos de tipo atributo o cualitativos: Expresan una cualidad de un individuo. En R guardaremos als listas de datos cualitativos en vectores (habitualmente de palabras) o en factores si vamos a usarlos para clasificar individuos.

- Datos ordinales: Similares a los cualitativos, con la unica diferencia de que se pueden ordenar de manera natural. Por ejemplo, las calificaciones en un control (suspendido, aprobado, notable, sobresaliente). En R guardaremos las listas de datos ordinales en factores ordenados.
- Datos cuantitativos: Se refieren a medidas, tales como edades, longitudes, etc. En R guardaremos las listas de datos cuantitativos en vectores numéricos.

  Para datos cualitativos tiene sentido identificar frecuencias y demás valores. mientrsas que para datos cuantitativos tiene sentido involucrar media, desv estandar, mediana, etc.

```
str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Como podemos ver en la anterior tenemos 4 variables cuantitativas y una variable cualitativa.

DATOS CUALITATIVOS

Corresponden a observaciones sobre cualidades de un objeto o individuo. Suelen codificarse por medio de palabras, pero también se pueden usar números que jueguen el papel de etiquetas.

> Ejemplo: es habitual representar No (o falso,fracaso,ausente,etc.) con un 0 y si (o verdadero, exito, presente, etc) con un 1.

En general los datos cualitativos son aquellos que pueden ser iguales o diferentes pero que no admiten ningun otro tipo de comparación significativa.

Es decir, que no tenga ningun sentiudo preguntarse si uno es más grande que otro, ni efectuar operaciones aritméticas con ellos, aunque estén representados por números.

'¿Qué son los datos cualitativos?

Por lo tanto, un mismo coonjutno de datos puede ser cualitativo o de otro tipo, según el análisis que vayamos a hacer de el.

> Ejemplo: Si hemos anotado durante unos años los dias de la semana en los que ha llovido y queremos contar cuantas veces ha ocurrido el lunes, cuantas en martes, etc,.. esta lsita de nombres o numeros serán datos cualitativos(llover o no llover). Si en cambio, queremos estudiar como se comportan los dias de lluvia según avanza la semana y por lo tanto el oden de los dias es relevante serán ordinales (**DESDE EL MOMENTO EN QUE TE INTRERESA UN ORDEN YA DEJA DE SER UN DATO CUALITATIVO SINO ORDINAL**).

VARIABLE CUALITATIVA

Lista de observaciones de un tipo de datos cualitativos sobre un conjunto concreto de objetos.

NIVELES

Diferentes valores que pueden tomar estos datos. Por ejemplo, los dos niveles de la variable Sexo, serían M (Macho) H(Hembra), o sinonimos.

Con R usaremos vectores, y factores para representar variables cualitativas. Los factores nos servirpan para agrupar las observaciones según los niveles de la variable. De esta manera podremos segmentar la población

que representa la variable en grupos o subpoblaciones, asignando un grupo a cada nivel, y podremos comparar el comportamiento de otras variables sobre estos grupos.

ESTUDIO DE FRECUENCIAS

Las estadísticas básicas que se pueden sacar para datos cualitativos son:

- Contar cuántos datos hay en ese nivel **frecuencia absoluta**
- Qué fracción del total representan **frecuencia relativa**

Ejemplo:

Supongamos que tenemos un tipo de datos cualitativos con niveles:

$$l_1, l_2, ..., l_k$$

Efectuamos n observaciones de este tipo de datos, y denotamos por:

$$x_1, x_2, ..., x_n$$

Los resutlados que obtenemos con:

$$x_j \in \{l_1, l_2, ..., l_k\}$$

Con estas notaciones:

- La **frecuencia absoluta**, n_j, del nivel l_j en esta variable cualitativa es el numero de observaciones en las que x_i toma el valor l_j.
- La **frecuencia relativa** del nivel l_j en esta variable es la fracción:

$$f_i = \frac{n_j}{n}$$

  Es decir, la frecuencia relativa del nivel l_j es la fracción (en tanto por uno) de observaciones que corresponden a este nivel.

La **moda** de esta variable cualitativa es su nivel, o niveles de mayor frecuencia (absoluta o relativa).

EJEMPLO

Ejemplo: Supongamos que se ha realizado un seguimietno a 20 personas asistentes a un congreso. Uno de los datos que se han recogido sobre las 20 personas ha sido su sexo. El resultado ha sido una variable cualitativa dformada por las 20 observaciones siguientes:

Mujer,Mujer,Hombre,Mujer,Mujer,Mujer,Mujer,Mujer,Hombre,Mujer,Hombre,Hombre,Mujer,Mujer,Hombre,Mujer,Mujer,Muje

Sus dos niveles son Hombre y Mujer. En esta variable hay 14 mujeres y 6 hombres. Éstas son las frecuencias absolutas de estos niveles.

Puesto que hay en total 20 individuos, sus frecuencias relativas son:

$$Hombre = \frac{6}{20} = \ 0.3$$
$$Mujer = \frac{14}{20} = \ 0.7$$

En este caso $l_1 = Hombre$ y $l_2 = Mujer$, $n = 20$ (el numero de observaciones efectuadas), y $x_1, ..., x_{20}$ formarían la muestra de los sexos.

Normalmente cuando tenemos ese tipo de informacion se suele representar en formato de tabla:

Sexo	n_i	f_i	%
Hombre	6	0.3	30%
Mujer	14	0.7	70%
Total	20	1	100%

Su moda es el nivel **Mujer**

TABLAS DE FRECUENCIAS UNIDIMENSIONALES

Supongamos que tenemos una variable cualitativa guardada en un vector o un factor como la siguiente:

```
x=sample(1:5,size = 12,replace = TRUE)#rango de valores,tamaño, siempre lleva replace TRUE dado que el
x
```

```
 [1] 1 3 3 4 4 5 1 2 3 3 2 2
```

```
respuestas=factor(sample(c("Si","No"),size = 12,replace = TRUE))
respuestas
```

```
 [1] Si Si Si No No No Si No No Si No No
Levels: No Si
```

Con R podemos utilizar la funcion table para sacar la tabla de frecuencias absolutas automáticamente.

```
table(x)
```

```
x
1 2 3 4 5
2 3 4 2 1
```

```
table(respuestas)
```

```
respuestas
No Si
 7  5
```

Que en el muno de la estadística se comoce como **tabla de contingencia**, la primer fila son los niveles o valores y la fila de abajo son el conteo.

TABLAS DE FRECUENCIA UNIDIMENSIONALES

El resultado de una funcion **table()** es un objeto de datos de un tipo nuevo: una tabla de contingencia, una table en el argot de R.

Al aplicar table() a un vector obtenemos una tabla unidimensional formada por una fila con los niveles de la variable y una segunda fila donde, debajo de cada nivel, aparece su frecuencia absoluta en el vector.

Los nombres de las columnas de una tabla unidimensional se obtienen con la funcion **names()**.

```
names(table(x))
```

```
[1] "1" "2" "3" "4" "5"
```

```
names(table(respuestas))
```

```
[1] "No" "Si"
```

En la **table** de un vector sólo aparecen los nombres de los niveles presentes en el vector. Si el tipo de datos cualitativos usado tenía más niveles y queremos que aparezcan explícitamente en la tabla (con frecuencia 0), hay que transformar el vector en un factor con los niveles deseados.

```
z=factor(x,levels=1:7)
z
```

```
 [1] 1 3 3 4 4 5 1 2 3 3 2 2
Levels: 1 2 3 4 5 6 7
```

Lo anterior se usa si se quiere incluir dentro de los niveles del factor aunque no estén en el original. es decir.

```
names(table(z))
```

```
[1] "1" "2" "3" "4" "5" "6" "7"
```

```
table(z)
```

```
z
1 2 3 4 5 6 7
2 3 4 2 1 0 0
```

Podemos pensar que una tabla unidimensional es como un vector de numeros donde cada entrada está identificada por un nombre: el de su columna. Para referirnos a una entrada de una tabla unidimensional, podemos usar tanto su posición como su nombre (entre comillas, aunque sea un número).

```
table(x)[3]# la columna 3 de table(x)
```

```
3
4
```

```
table(x)["7"]# la columna de table(x) con nombre 7
```

```
<NA>
  NA
```

```
table(x)["5"]# la columna de table(x) con nombre 5
```

```
5
1
```

```
table(x)[2]
```

```
2
3
```

```
3*table(x)[2]# el triple de la segunda columna de table(x)
```

```
2
9
```

Aplicamos la tabla de contingencia aceptan la mayoría de las funciones que ya hemos utilizado para vectores.

```
table(x)
```

```
x
1 2 3 4 5
2 3 4 2 1
```

```
sum(table(x))#suma las entradas de table x, es decir el conteo
```

```
[1] 12
```

```
sqrt(table(respuestas))#raices cuadradas de las entradas de table(Respuestas)
```

```
respuestas
      No       Si
2.645751 2.236068
```

Por ejemplo

```
dale=factor(sample(c("H","M"),size = 10,replace = TRUE))
dale
```

```
 [1] H M M M H H H M H H
Levels: H M
```

```
table(dale)
```

```
dale
H M
6 4
```

```
table(dale)["M"]
```

```
M
4
```

```
sum(table(dale))# retorna el total de observaciones vistas
```

```
[1] 10
```

TABLA DE FRECUENCIAS RELATIVAS

La tabla de frecuencias relativas de un vector se puede calcular aplicando la funcion **prop.table() a su table**. El resultado vuelve a ser una tabla de contingencia unidimensional.

```
prop.table(table(x))
```

```
x
         1          2          3          4          5
0.16666667 0.25000000 0.33333333 0.16666667 0.08333333
```

```
prop.table(table(respuestas))
```

```
respuestas
       No        Si
0.5833333 0.4166667
```

Es importante aplicar la funcion prop.table() directamente al resultado de table por que sino salta error en caso de que sea vectore de palabras, en caso contrario de que sea numeros dará una tabla de frecuencias relativas de una variable que tuviera como tabla de frecuencias absolutas este vectore de numeros, entendiendo que cada entrada del vector representa la frecuencia de un nivel diferente.

```
x=c(1,1,1)
prop.table(table(x))
```

```
x
1
1
```

```
prop.table(x)
```

```
[1] 0.3333333 0.3333333 0.3333333
```

Tambien se puede calcular la tabla de frecuencias relaticvas de un vector dividiendo el resultado de table por el numero de observaciones

```
table(x)/length(x)# valor de la columna entre la longitud total de observaciones
```

```
x
```

```
1
1
```

Si queremos obtener dado un vector x el nivel que se repita exactamente n veces debemos usar la siguiente instrucción:

```
x=c(1,1,2,2,2,2,4,4,4,5,5,5,5,5)
names(which(table(x)==1))# aquellos niveles que se repitan una vez
```

```
character(0)
```

```
names(which(table(x)==3))# aquellos niveles que se repitan 3 veces
```

```
[1] "4"
```

Podemos usar diferentes condicionales.

De esa manera podemos calcular la moda con la instrucción anterior quedando de la siguiente manera:

```
names(which(table(x)==max(table(x))))#moda del vector y el out es el nivel
```

```
[1] "5"
```

```
names(which(table(respuestas)==max(table(respuestas))))
```

```
[1] "No"
```

EJERCICIO

```
hom=c(rep("H",6))
muj=c(rep("M",14))
dft=c(hom,muj)
dft=factor(dft)
#CALCULO DE TABLA DE FRECUENCIA ABSOLUTA

table(dft)
```

```
dft
 H  M
 6 14
```

```
##CALCULO DE TABLA DE FRECUENCIA RELATIVA
prop.table(table(dft))
```

```
dft
  H   M
0.3 0.7
```

```
#CALCULO DE MODA
names(which(table(dft)==max(table(dft))))
```

```
[1] "M"
```

```
funModa<-function(tabla){
  names(which(table(tabla)==max(table(tabla))))
}
m_t=funModa(dft)
```

La moda del dataframe es M‘

TABLAS DE FRECUENCIAS BIDIMENSIONALES

La función table() permite construir tablas de frecuencias conjuntas de dos o más variables.

Supongamso que el vector **respuestas** anterior contiene las respuestas a una pregunta dadas por unos individuos cuyos sexos tenemos almacenados en un vector **sexo**, en el mismo orden que sus respuestas. En este caso, podemos construir una tabla que nos diga cuántas personas de cada sexo han dado respuesta.

```
sexo=sample(c("H","M"),size = length(respuestas),replace = TRUE)
table(respuestas,sexo)
```

```
          sexo
respuestas H M
        No 4 3
        Si 3 2
```

```
table(sexo,respuestas)
```

```
    respuestas
sexo No Si
   H  4  3
   M  3  2
```

```
t(table(sexo,respuestas))#transpuesta
```

```
          sexo
respuestas H M
        No 4 3
        Si 3 2
```

Es muy cómodo utilizar las respuestas en columnas y los niveles en las filas.

Para referirnos a una entrada de una tabla bidimensional podemos usar el sufijo [,] como si estuvieramos en una matriz o dataframe. Dentro de los corchetes, tanto podemos usar los indices como los nombres entre comillas de los niveles.

```
table(respuestas,sexo)["Si","H"]
```

```
[1] 3
```

```
table(respuestas,sexo)[1,1]
```

```
[1] 4
```

```
  table(respuestas,sexo)[2,2]
```

```
[1] 2
```

La funcion **prop.table()** sirve para calcular tablas bidimensionales de frecuencias relativas conjuntas de pares de variables. Pero en el caso bidimensional tenemos dos tipos de frecuencias relativas:

- Frecuencias relativas globales: para cada par de niveles, uno de cada variable, la fracción de individuos que pertenecen a ambos niveles respecto del total de la muestra.
- Frecuencias relativas marginales: Dentro de cada nivel de una variable y para cada nivel de la otra, la fracción de individuos pertenecen al segundo nivel respecto del total de la subpoblación definida por el primer nivel.

```
prop.table(table(sexo,respuestas))
```

```
    respuestas
sexo        No        Si
```

```
   H 0.3333333 0.2500000
   M 0.2500000 0.1666667
```

De este modo, la tabla prop.table(table(sexo,respuestas)) nos da la fracción del total que representa cada pareja (sexo,respuesta)

Para poder obtener las tablas de frecuencias relativas marginales es necesario aplicar el siguioente comando:

Para obtener las marginales, debemos usar el parámetro margin al aplicar la funcion prop.table() a la table. Con margin=1 obtenemos las frecuencias relativas de las filas y con margin=2, de las columnas.

```
prop.table(table(sexo,respuestas),margin = 1)#por sexo
```

```
    respuestas
sexo        No        Si
   H 0.5714286 0.4285714
   M 0.6000000 0.4000000
```

```
#del total de hombres 20 porciento respondieron que no y 80 que si
#del taotal de mujeres 29 respondieron que no y 71 que si
prop.table(table(sexo,respuestas),margin=2)#por respuestas
```

```
    respuestas
sexo        No        Si
   H 0.5714286 0.6000000
   M 0.4285714 0.4000000
```

```
#del total de personas que dijeron no 33 pórciento son hombres y 66 mujeres
#del total de sis 44 son hombres y 55 son mujeres
```

> **MARGINAL** ES LO MISMO A FRECUENCIAS ABSOLUTAS es el conteo de cuantos datos hay para ese nivel, la frecuencia **RELATIVA** hace referencia a la fraccion respecto del total que representan.
>
> La relativa global se mide con respecto el total de la población y la relativa marginal fijan para cada fila o columna como se distribuyen
>
> La tabla de **CONTINGENCIA** es la de frecuancias absolutas

LA FUNCIÓN CROSS TABLE

La función **CrossTable()** Viene en el paquete **gmodels** permite producir (especificando el parámetro prop.chisq=FALSE) un resumen de la tabla de frecuencias absolutas y las tres tablas de frecuencias relativas de dos variables en un formato adecuado para su visualización.

La leyenda Cell Contents explica los contenidos de cada tabla: la frecuencia absoluta, la frecuencia relativa por filas, la frecuencia relativa por columnas y la frecuencia relativa global. Esta función dispone de muchos parámetros que permiten modificar el contenido de las celdas y que podemos consultar en help(CrossTable).

```
library(gmodels)

sexo=factor(sample(c("H","M"),size = 10,replace = TRUE))
ans=factor(sample(c("S","N"),size = 10,replace = TRUE))
CrossTable(sexo,ans,prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
```

```
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  10
##
##
##              | ans
##         sexo |         N |         S | Row Total |
## -------------|-----------|-----------|-----------|
##            H |         3 |         3 |         6 |
##              |     0.500 |     0.500 |     0.600 |
##              |     0.600 |     0.600 |           |
##              |     0.300 |     0.300 |           |
## -------------|-----------|-----------|-----------|
##            M |         2 |         2 |         4 |
##              |     0.500 |     0.500 |     0.400 |
##              |     0.400 |     0.400 |           |
##              |     0.200 |     0.200 |           |
## -------------|-----------|-----------|-----------|
## Column Total |         5 |         5 |        10 |
##              |     0.500 |     0.500 |           |
## -------------|-----------|-----------|-----------|
##
##
```