Exploración de datos y estadística descriptiva

Christian Salas Eljatib, Ph.D.

E-mail: cseljatib@gmail.com

Web: https://eljatib.com

4 de enero de 2023
Santiago, Chile

Contenidos

Dataframe: idahohd (desde paquetes datana)

Dataframe: idahohd

```
> library(datana)
> data(idahohd)
```

(*) En caso que no tenga instalada el paquete, carga el archivo de datos `ufcBiometria.csv`, como hicimos antes

```
> idahohd <- read.csv("ufcBiometria.csv")
```

Revisando

```
> dim(idahohd) #dimensiones (num filas, num columnas)
[1] 372   5
> names(idahohd) #nombre de las variables
[1] "plot"    "tree"    "species" "dbh"     "height"
> head(idahohd)
  plot tree species dbh height
1    2    1      DF 390    205
2    2    2      WL 480    330
3    3    2      GF 520    300
4    3    5      WC 360    207
5    3    8      WC 380    225
6    4    1      WC 460    180
```

prueba también

```
> idahohd[100:103, ]
> idahohd[100:103, c("dbh","height")]
```

Cambiando el nombre de las variables (columnas)

```
> names(idahohd) #primero veamos el nombre de las variables
[1] "plot"    "tree"    "species" "dbh"     "height"
> # y ahora cambiemos sus nombres
> names(idahohd) <- c("parce","arbol","spp","dap","altura")
> # revisemos si esta OK ahora
> names(idahohd)
[1] "parce"  "arbol"  "spp"    "dap"    "altura"
```

Consejos

- No utilizar tildes ni espacios para nombrar archivos
- Evitar emplear mayúsculas para nombrar variables, mantener todo en minúsculas.

Creando nuevas variables

```
> idahohd$d <- idahohd$dap/10
> idahohd$h <- idahohd$altura/10

> idahohd$ln.d <- log(idahohd$d)
> idahohd$ln.h <- log(idahohd$h)
```

Ahora veamos nuestros datos luego de haber realizado algunos cambios

```
> head(idahohd)
  parce arbol spp dap altura  d    h   ln.d   ln.h
1     2     1  DF 390    205 39 20.5 3.6636 3.0204
2     2     2  WL 480    330 48 33.0 3.8712 3.4965
3     3     2  GF 520    300 52 30.0 3.9512 3.4012
4     3     5  WC 360    207 36 20.7 3.5835 3.0301
5     3     8  WC 380    225 38 22.5 3.6376 3.1135
6     4     1  WC 460    180 46 18.0 3.8286 2.8904
```

Seleccionando variables

```
> idahohd <- idahohd[,c('parce','arbol','spp','d','h', "ln.d", "ln
> ncol(idahohd) #el numero de columnas ahora disminuyo
[1] 7
```

Guardando una dataframe a un archivo

```
> write.csv(idahohd, "newIdahohd.csv")
```

Seleccionar una porción de un dataframe

```
> class(idahohd$spp) #Que clase de variable es?
[1] "character"
> unique(idahohd$spp)
 [1] "DF" "WL" "GF" "WC" "WP" "SF" "LP" "HW" "ES" "PP"


> idahohdGF <- subset(idahohd,spp=="GF")


> nrow(idahohd) #numero de filas de la dataframe original
[1] 372
> nrow(idahohdGF) #numero de filas de la dataframe solo con G
[1] 111
```

Subdivisión en base a una variable continua

Según el valor de una variable

Empleemos como referencia que la variable diámetro del tronco (d) sea mayor a 95 cm

```
> subset(idahohd, d > 95)
    parce arbol spp     d    h   ln.d   ln.h
132    43     4  WC 101.5 39.0 4.6201 3.6636
141    44     3  WP 112.0 38.5 4.7185 3.6507
173    55     2  DF  99.8 42.0 4.6032 3.7377
230    78     3  WP 103.0 48.0 4.6347 3.8712
```

Subdivisión en base a una variable categórica

Sólo una especie

Empleemos como referencia a "HW", nombre común: western hemlock, nombre científico: *Tsuga heterophylla*

```
> subset(idahohd, spp == 'HW')
    parce arbol spp    d  h   ln.d   ln.h
73     23     4  HW 20.7 20 3.0301 2.9957
182    57     5  HW 31.2 24 3.4404 3.1781
187    59     6  HW 17.6 18 2.8679 2.8904
188    59     7  HW 17.5 19 2.8622 2.9444
```

Subdivisión en base a más de una característica

operación lógica **y**

Según el valor de una variable continua **y** de una variable categórica

Empleemos como referencia que la variable diámetro del tronco (d) sea mayor a 95 cm **y** que la especie sea "DF", nombre común: Douglas-fir, nombre científico: *Pseudotsuga menziesii*

```
> subset(idahohd, d > 95 & spp=="DF")
    parce arbol spp    d  h   ln.d   ln.h
173    55     2  DF 99.8 42 4.6032 3.7377
```

Subdivisión en base a más de una característica

operación lógica **o**

Según el valor de una variable continua **o** de una variable categórica

Empleemos como referencia que la variable diámetro del tronco (d) sea mayor a 95 cm **o** que la especie sea "PP", nombre común: Pino ponderosa, nombre científico: *Pinus ponderosa*

```
> subset(idahohd, d > 95 | spp=="PP")
    parce arbol spp     d    h   ln.d   ln.h
132    43     4  WC 101.5 39.0 4.6201 3.6636
141    44     3  WP 112.0 38.5 4.7185 3.6507
173    55     2  DF  99.8 42.0 4.6032 3.7377
185    59     2  PP  76.9 40.0 4.3425 3.6889
230    78     3  WP 103.0 48.0 4.6347 3.8712
345   128     7  PP  37.1 26.0 3.6136 3.2581
```

Estadísticos de posición

- **Media aritmética**

$$\overline{y} = \frac{1}{n}\sum_{i}^{n} y_i \tag{1}$$

- **Media geométrica**

$$y_{GM} = \left(\prod_{i=1}^{n} y_i\right)^{1/n} = \sqrt[n]{\prod_{i}^{n} y_i} \tag{2}$$

- **Mediana** El estadístico que separa en dos la distribución de una variable de una muestra, población, o pdf.
  Esto es, la mediana y_M cumple con la siguiente relación

$$0{,}5 = \int_{-\infty}^{y_M} f(y)dy, \text{ o bien} \tag{3}$$

$$0{,}5 = \int_{y_M}^{\infty} f(y)dy \tag{4}$$

Cuantiles (Percentiles)

Suponiendo que hacemos que la integral de una pdf, o como la denominamos $F(y)$, sea igual a un valor arbitrario p, cumpliendo entonces

$$F_y(Y) = \Pr(Y \leq y) = p, \tag{5}$$

entonces y se denomina el p-ésimo **cuantil** ["quantile"] de la distribución F, como $\mathcal{Q}^{(p)}(F)$ o simplemente $\mathcal{Q}^{(p)}$.
También puede ser y_p

Cuartiles

Cuartil 1 = $Q1$ = percentil 25
Cuartil 2 = $Q2$ = percentil 50
Cuartil 3 = $Q3$ = percentil 75

Deciles y quintiles

Deciles: Dividen la pdf en **10** partes iguales
Quintiles: Dividen la pdf en **5** partes iguales

Estadísticos de dispersión

- **Desviación estándar**

$$\widehat{\sigma}_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = s_y. \tag{6}$$

- **Rango**

$$\text{Rango} = y_{max} - y_{min} \tag{7}$$

- **Diferencia absoluta media**

$$\text{DAM} = \frac{1}{n} \mid y_i - \overline{y} \mid \tag{8}$$

Note que también podríamos ocupar y_M en reemplazo de $\overline{y}$, o cualquier otra medida de tendencia central.

Estadísticos de dispersión (cont)

- **Coeficiente de variación**

$$CV_y = 100\frac{\widehat{\sigma}_y}{\overline{y}} \tag{9}$$

- **Rango intercuartílico**

$$\text{RIQ} = Q3 - Q1 = y_{75} - y_{25} \tag{10}$$

Estadística descriptiva en R

summary() para una variable

```
> summary(idahohd$d)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   10.0    23.0    33.8    36.6    46.4   112.0
```

summary() para más de una variable

```
> summary(idahohd[,c('d','h","ln.d","ln.h")])
       d                h               ln.d            ln.h
 Min.   : 10.0   Min.   : 5.0   Min.   :2.30   Min.   :1.61
 1st Qu.: 23.0   1st Qu.:19.0   1st Qu.:3.13   1st Qu.:2.94
 Median : 33.8   Median :24.5   Median :3.52   Median :3.20
 Mean   : 36.6   Mean   :24.3   Mean   :3.48   Mean   :3.14
 3rd Qu.: 46.4   3rd Qu.:29.5   3rd Qu.:3.84   3rd Qu.:3.38
 Max.   :112.0   Max.   :48.0   Max.   :4.72   Max.   :3.87
```

Estadística descriptiva (cont.)

Mínimo, máximo, media, desv. estándar [min(), max(), mean(), sd()]

```
> mean(idahohd$h)
[1] 24.313
> sd(idahohd$h)
[1] 7.3824
```

Percentiles [quantile(x,p)], e.g., quantile(x,0.5) = median(x)

```
> quantile(idahohd$d,0.25)
   25%
22.975
> quantile(idahohd$d,c(0.25,0.6,0.9))
   25%    60%    90%
22.975 39.180 59.360
```

Estadística descriptiva por niveles de factores

tapply() y mean()

```
> tapply(idahohd$d,idahohd$spp,mean)
    DF     ES     GF     HW     LP     PP     SF     WC     WL     WP
39.764 46.500 35.343 21.750 24.733 57.000 13.550 39.282 31.875 35.590
```

tapply() y sd()

```
> tapply(idahohd$d,idahohd$spp,sd)
     DF      ES      GF      HW      LP      PP      SF      WC
16.5534 12.0208 17.1129  6.4728  5.9341 28.1428  3.1592 18.3913
     WL      WP
13.7256 23.1700
```

Tamaño muestral

```
> tapply(idahohd$d,idahohd$spp,length)
 DF  ES  GF  HW  LP  PP  SF  WC  WL  WP
 55   2 111   4   3   2  10 136  20  29
```

Pruebe además con

```
> tapply(idahohd$d,idahohd$spp,summary)
```

Tabla de estadística descriptiva Función: descstat()

```
> names(idahohd)
[1] "parce" "arbol" "spp"   "d"     "h"     "ln.d"  "ln.h"
> db.aqui <- idahohd[,c('d','h','ln.d','ln.h')]


> descstat(db.aqui,1) #un numero decimal para el output
                d     h  ln.d  ln.h
n           372.0 372.0 372.0 372.0
Minimum      10.0   5.0   2.3   1.6
Maximum     112.0  48.0   4.7   3.9
Mean         36.6  24.3   3.5   3.1
Median       33.8  24.5   3.5   3.2
Std. Dev.    18.2   7.4   0.5   0.3
CV %         49.6  30.4  14.7  10.7
```