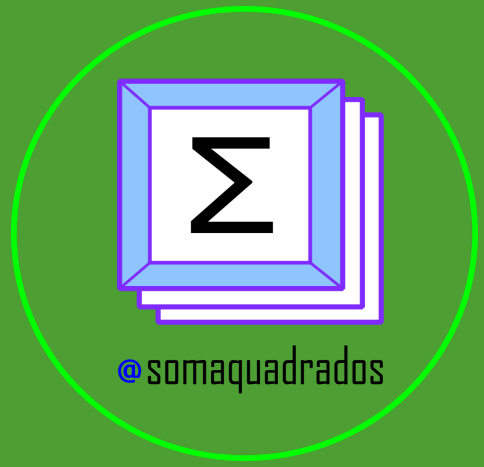
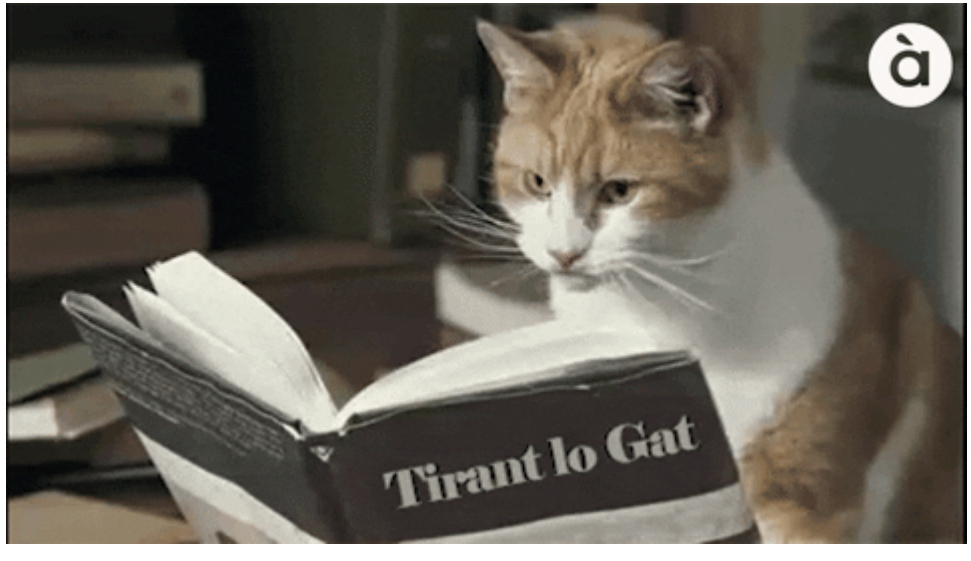
Introducción al análisis de datos biológicos con R



Eliana F. Burgos

Contenidos

@somaquadrados

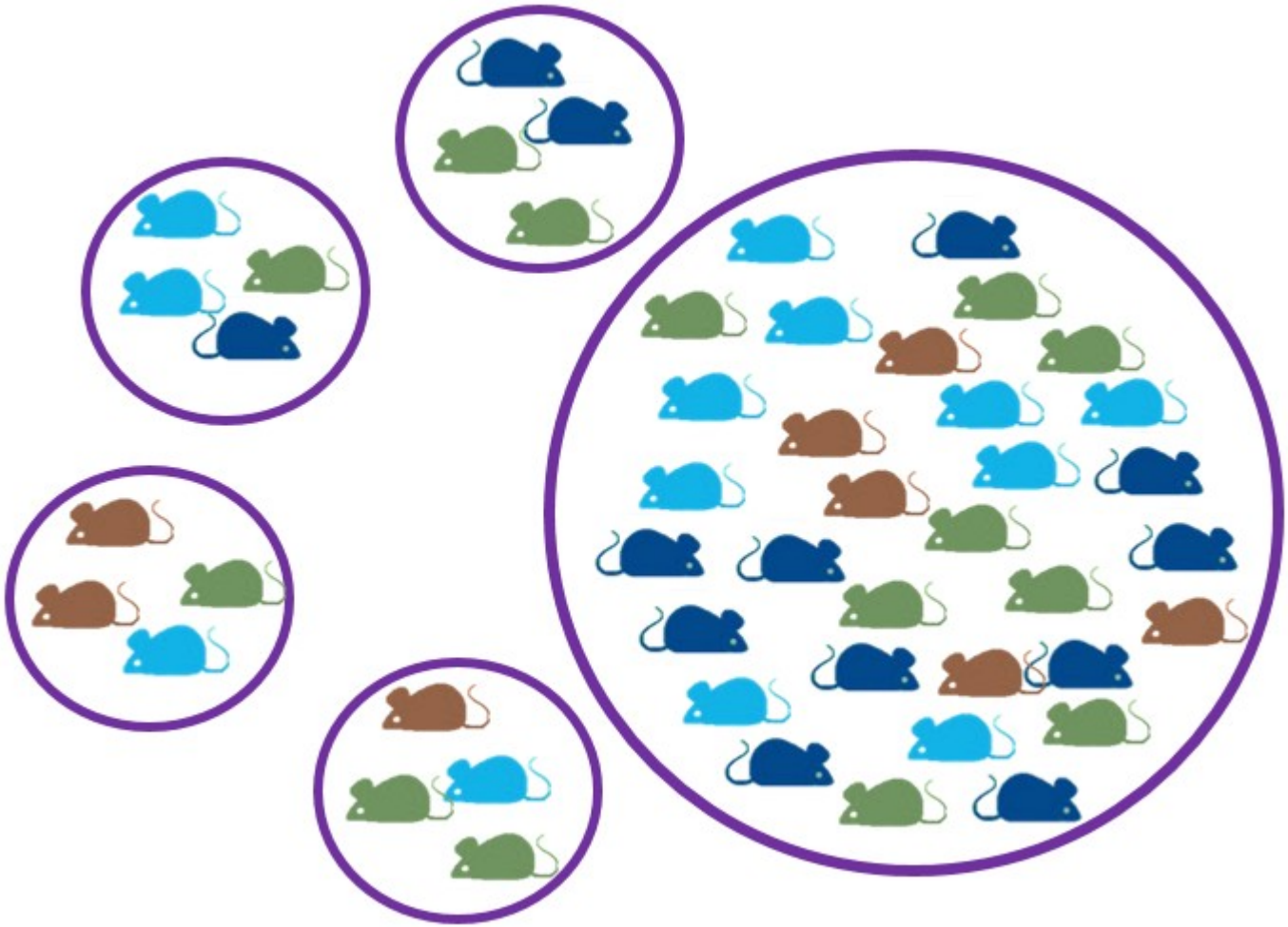


Variables estadísticas

Variables estadísticas

@somaquadrados

Elementos de la población -> unidades estadísticas



Variables estadísticas

@somaquadrados

- **Variables cualitativas**
  - *ordinales*
  - *nominales*
- **Variables cuantitativas**
  - *discretas*
  - *continuas*

Variables estadísticas

@somaquadrados

- **Variables cualitativas**
  - *ordinales* **->** tamaño (grande, mediano, pequeño); distancia (muy lejos, lejos, cerca); clases de edades; año de muestreo; tratamiento (1,2,3;a,b,c)
  - *nominales* **->** sexo, especie, sitio, color, uso del suelo, cuadrante, presencia/ausencia
- **Variables cuantitativas**
  - *discretas* **->** abundancia, indv. positivos, cantidad de huevos/embriones
  - *continuas* **->** índices, peso, largo, temperatura, humedad, intensidad luminica

Ejemplo

@somaquadrados

Contents lists available at ScienceDirect

Mammalian Biology

journal homepage: www.elsevier.de/mambio

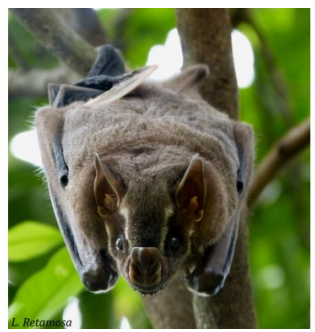
Original Investigation

Bat frugivory in two subtropical rain forests of Northern Argentina: Testing hypotheses of fruit selection in the Neotropics

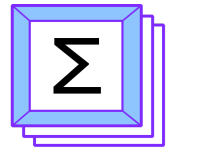
Mariano S. Sánchez[a,b,*], Norberto P. Giannini[a,b,c], Rubén M. Barquez[a,b]

[a] Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
[b] Programa de Investigaciones de Biodiversidad Argentina, Facultad de Ciencias Naturales e Instituto Miguel Lillo, Universidad Nacional de Tucumán, Miguel Lillo 205, C. P. 4000, Tucumán, Argentina
[c] Department of Mammalogy, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024-5192, USA

OBJETIVO: evaluar la dieta y el nicho dietario de especies de murciélagos frugívoros.



Ejemplo



METODOLOGíA

Datos: recopilaron datos novedosos y utilizaron datos de estudios previos

- **Murcielagos**
  - masa corporal
  - largo del antebrazo
  - sexo (macho, hembra)
  - edad (juvenil, adulto)
- **Vegetación**: colectaron ejemplares en cada sitio
  - tamaño individual de las semillas
  - color de la fruta
  - forma de la fruta
  - n° de semillas por fruto
  - habitat (bosque primario vs bosque secundario ribereño)

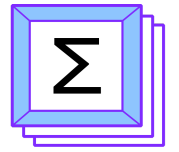
@somaquadrados

Parámetro	Tipo de variable
Murcielagos	
masa corporal	cuantitativa continua
largo del antebrazo	cuantitativa continua
sexo	cualitativa nominal de dos niveles: macho/hembra
edad	cualitativa ordinal de dos niveles: juvenil/adulto

@somaquadrados

Parámetro	Tipo de variable
Vegetación	
tamaño de semilla	cuantitativa continua
color del fruto	cualitativa nominal
forma del fruto	cualitativa nominal
n° de semillas por fruto	cuantitativa discreta
habitat	cualitativa nominal de dos niveles: bosque primario/bosque secundario

IMPORTANTE



Cuando cargamos la planilla al R, el programa lee cada columna con la categoria asignada en el procesador de bases de datos

No siempre esa categoria es la correcta y hay que configurarla

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   fecha = col_character(),
##   sitio = col_character(),
##   uso_suelo = col_character(),
##   ambiente = col_character(),
##   estacion = col_character(),
##   anio.estacion = col_character()
## )
## i Use `spec()` for the full column specifications.
```

Entonces

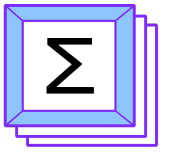
variable cualitativa


```
class(datos$uso_suelo)
```

```
## [1] "character"
```

```
datos$uso_suelo <- as.factor(datos$uso_suelo)
class(datos$uso_suelo)
```

```
## [1] "factor"
```



variable numérica continua

```
datos$IDR_total <- as.numeric(datos$IDR_total)
class(datos$IDR_total)
```

```
## [1] "numeric"
```

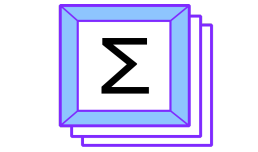
variable numérica discreta

```
class(datos$n_total)
```

```
## [1] "numeric"
```

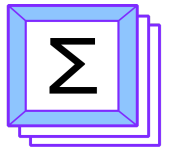
```
datos$n_total <- as.integer(datos$n_total)
```

Estadistica descriptiva



- *descripción de una población*
- descripción de las diferentes variables
- teniendo en cuenta:
  - valor medio
  - dispersión/variación
  - forma

Medidas de posición



@somaquadrados

- son medidas de tendencia central
- marcan la acumulación de los datos en torno a un valor
- **media, mediana y moda**

Media

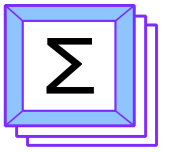
- muestra el valor promedio de nuestra variable de interés

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Ejemplo

Estamos estudiando el estado de conservación de la especie *Chironectes minimus* mejor conocida como **cuica de agua**, una zarigüeya propia de la Selva Paranaense. Realizamos dos muestreos por año (primavera-verano; otoño-invierno) entre los años 2010 al 2012 en áreas naturales protegidas y en cultivos forestales y de yerba donde existen cuerpos de agua (lagunas y arroyos).

Objetivo 1: Evaluar los cambios en la abundancia de esta especie en los diferentes ambientes estudiados.



Media de toda la muestra

```
mean(cuica$ncuicas, na.rm = TRUE)
```

```
## [1] 9.466667
```

```
round(mean(cuica$ncuicas, na.rm = TRUE))
```

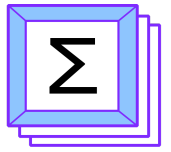
```
## [1] 9
```

Media de un conjunto de datos recortado

```
round(mean(cuica$ncuicas, na.rm = TRUE, trim = 0.10))
```

```
## [1] 9
```

El comando *trim()* nos permite indicar los datos que queremos excluir de cada extremo de la distribución



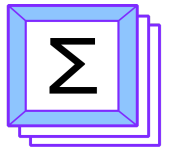
@somaquadrados

Media de un conjunto de datos que cumplen un criterio

```
library(tidyverse)
f1 <- filter(cuica, ambiente=="yerba")
round(mean(f1$ncuicas,na.rm = TRUE))
```

```
## [1] 6
```

Si además de los datos que tomamos en campo, contamos con estudios previos y queremos conocer la media de esa población utilizando todos los datos, se puede calcular la media ponderada



@somaquadrados

Esto se realiza con una adaptación de la formula donde incluimos la *media* y el *n* de las observaciones. Por ejemplo: nuestros datos tienen una media de 6 (*n*=10), y los estudios previos muestran medias de 15 (*n*=25), 7 (*n*=20) y 12 (*n*=18)

```
mpond <- ((6*10)+(15*25)+(7*20)+(12*18))/(10+25+20+18)
round(mpond)
```

```
## [1] 11
```

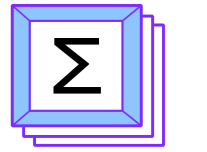
Mediana

- es el valor que se encuentra en la mitad en la lista ordenada de nuestros datos

```
median(cuica$ncuicas)
```

```
## [1] 8
```

Moda



- la moda es el valor más frecuente en nuestros datos

```
frecuencias <- data.frame(table(cuica$ncuicas))
moda <- frecuencias[which.max(frecuencias$Freq),1]
moda
```

```
## [1] 7
## Levels: 2 5 6 7 8 10 11 12 13 16 18 19
```

```
library(modeest)
mfv(cuica$ncuicas)
```

```
## [1] 7
```

Cuartiles, deciles y percentiles

- puntos tomados a intervalos regulares de la función de distribución de una variable
- medidas de localización o posición no central
- se calcula con **quantile()** del paquete *stats*

Cuartiles

```
round(quantile(cuica$ncuicas, prob=seq(0, 1, 1/4)))
```

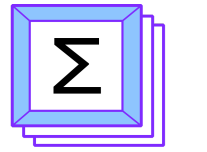
```
##   0%  25%  50%  75% 100%
##    2    6    8   12   19
```

Deciles

```
round(quantile(cuica$ncuicas, prob=seq(0, 1, length = 11)))
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    2    5    6    7    7    8   10   11   12   18   19
```

Percentiles




```
round(quantile(cuica$ncuicas,prob=seq(0, 1, length = 101)))
```

```
##   0%   1%   2%   3%   4%   5%   6%   7%   8%   9%  10%  11%  12%  13%  14%  15%
##    2    2    2    2    2    3    4    5    5    5    5    5    5    5    5    5
##  16%  17%  18%  19%  20%  21%  22%  23%  24%  25%  26%  27%  28%  29%  30%  31%
##    6    6    6    6    6    6    6    6    6    6    7    7    7    7    7    7
##  32%  33%  34%  35%  36%  37%  38%  39%  40%  41%  42%  43%  44%  45%  46%  47%
##    7    7    7    7    7    7    7    7    7    7    7    7    8    8    8    8
##  48%  49%  50%  51%  52%  53%  54%  55%  56%  57%  58%  59%  60%  61%  62%  63%
##    8    8    8    8    8    9    9   10   10   10   10   10   10   10   10   10
##  64%  65%  66%  67%  68%  69%  70%  71%  72%  73%  74%  75%  76%  77%  78%  79%
##   10   10   10   10   11   11   11   11   11   11   11   12   12   12   12   12
##  80%  81%  82%  83%  84%  85%  86%  87%  88%  89%  90%  91%  92%  93%  94%  95%
##   12   12   13   13   14   15   16   16   17   18   18   18   18   18   18   18
##  96%  97%  98%  99% 100%
##   18   18   18   19   19
```

Ejemplo

Abundancia de las zarigüeyas para el percentil 5, el 50 (mediana) y el 80.

```
round(quantile(cuica$ncuicas, probs=c(0.05, 0.5, 0.8)))
```

```
##  5% 50% 80%
##   3   8  12
```

Estadistica descriptiva

Medidas de dispersión

- nos permiten visualizar que tan variables o dispersos son nuestros datos
- valores mínimos y máximos, varianza, el desvío estandar, asimetria y curtosis, cuantiles

Mínimo y máximo

```
min(cuica$ncuicas)
```

```
## [1] 2
```

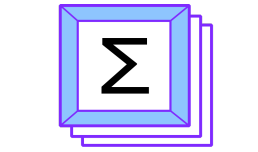
```
max(cuica$ncuicas)
```

```
## [1] 19
```

```
range(cuica$ncuicas)
```

```
## [1]  2 19
```

Varianza



- es el promedio de los cuadrados de los desvíos
- Es la esperanza del cuadrado de la desviación típica de dicha variable respecto a su media
- se expresa en la unidad de la variable al cuadrado

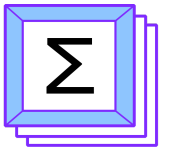
$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

```
round(var(cuica$ncuicas))
```

```
## [1] 22
```

```
round(sqrt((var(cuica$ncuicas))))
```

```
## [1] 5
```



@somaquadrados

Desvio estandar

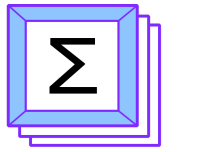
- es la raíz cuadrada de la varianza
- se expresa en la misma unidad en la que estan nuestros datos

$$s = \sqrt{s^2}$$

```
round(sd(cuica$ncuicas))
```

```
## [1] 5
```

Error estandar



- error estándar es la desviación estándar de la distribución muestral
- una estimación de la desviación estándar, derivada de una muestra particular usada para computar la estimación.
- es la desviación estándar dividida por la raíz cuadrada del número de observaciones.

```
library(plotrix)
round(std.error(cuica$ncuicas))
```

```
## [1] 1
```

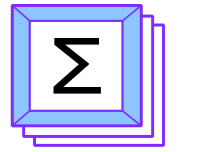
Cálculo manual

- EE es el desvío estandar dividido la raíz cuadrada del n° de observaciones

```
round(sd(cuica$ncuicas)/sqrt(length(cuica$ncuicas)))
```

```
## [1] 1
```

Coeficiente de variación



- Comparar dos grupos de datos de forma estandarizada
- permite comparar datos en diferentes escalas
- a >CV, menor representatividad de la media
- es una medida relativa

Datos que tomamos a campo

Teniamos una media de 9 con un desvío de 5

```
sd(cuica$ncuicas)/mean(cuica$ncuicas)
```

```
## [1] 0.4913594
```

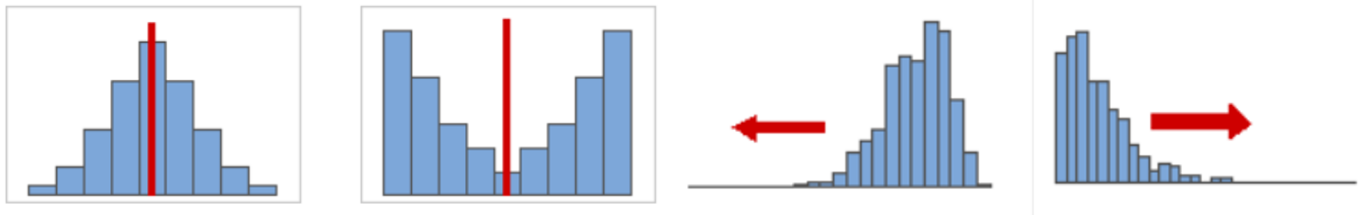
Asimetria y curtosis

@somaquadrados

- dan cuenta de la forma general de los datos
- nos permite identificar ciertas tendencias y comportamiento de los datos
- se utilizan comandos de la librería **psych**

Simetria y asimetria

- da cuenta de cómo se organizan los datos alrededor de la media
- pueden ser simétricas o asimétricas positivas o negativas
- los valor deben encontrarse entre -2 y 2.



Simetria y asimetria


```
skew(cuica$ncuicas)
```

```
## [1] 0.5910249
```

nos devuelve un valor en la escala de la variable que no nos permite comparar entre diferentes set de datos, entonces lo podemos estandarizar

```
skew(cuica$ncuicas)/sqrt(6/30)
```

```
## [1] 1.321572
```

Curtosis o apuntamiento

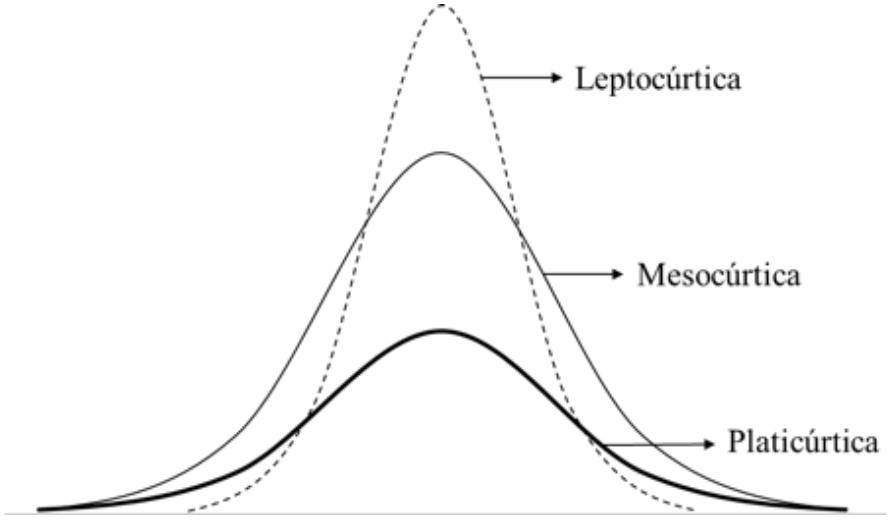
@somaquadrados

- mide que tan apuntada o achatada es la distribución de los datos al cercanos a la media

negativa la distribución es **platicúrtica**

igual a cero la distribución es **mesocúrtica**

positiva la distribución es **leptocúrtica**



Curtosis o apuntamiento

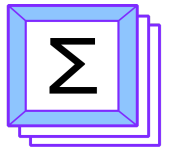

```
kurtosi(cuica$ncuicas)
```

```
## [1] -0.5590728
```

```
kurtosi(cuica$ncuicas)/sqrt(6/30)
```

```
## [1] -1.250125
```

¿Qué pasa si queremos comparar dos variables?



Covarianza

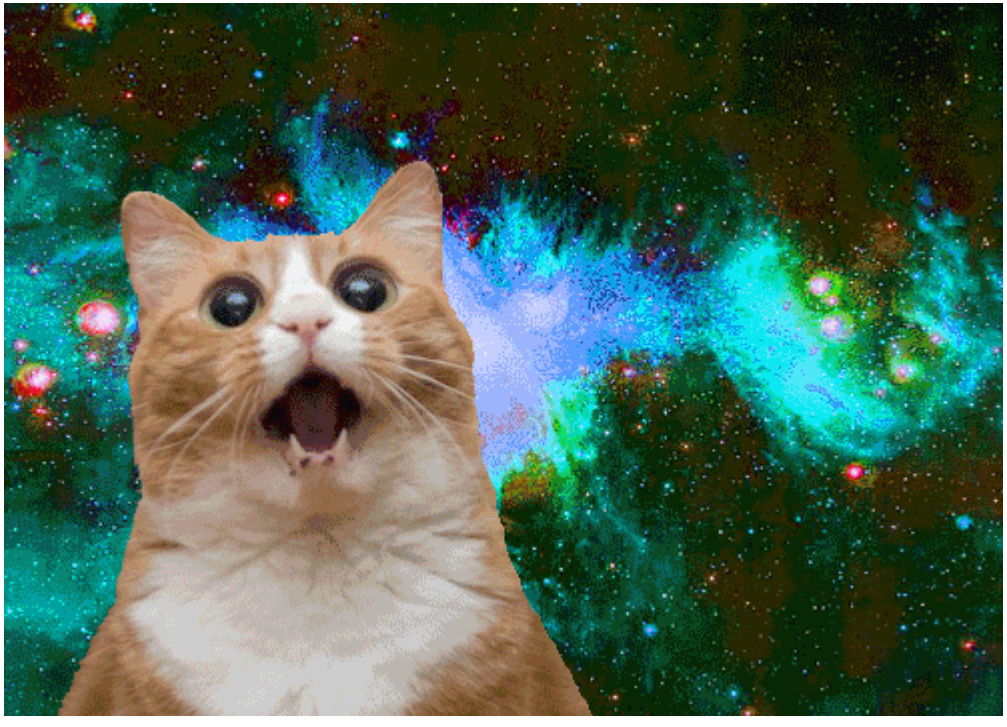
- mide la asociación lineal entre dos variables
- puede ser mayor, igual o menor que cero.
- será positica cuando la variable respuesta aumente con el aumento de la explicativa
- nos denota el tipo de relacion: positiva, negativa, neutra

```
cuica$dist_agua_m <- as.numeric(cuica$dist_agua_m)
cov(cuica$ncuicas, cuica$dist_agua_m)
```

```
## [1] -8.457471
```

Dudas y/o consultas

@somaquadrados



FIN