Curso Primeros Pasos en R

Clase 7: Análisis de datos usando R

Data Science UC

Pontificia Universidad Católica de Chile

Noviembre 2021

Clase 7: Análisis de datos usando R

- Introducción
- Algunos funciones estadísticas con `R Base`
- Resumen de datos con `dplyr`
- Tablas de frecuencia
- Análisis de correlación
- Taller práctico 1
- Taller práctico 2

Introducción

Conceptos básicos

Estadística descriptiva:

Está formada por los métodos gráficos y numéricos que se utilizan para recolectar, resumir y procesar los datos muestrales.

Estadística inferencial:

A partir de la descripción de datos muestrales y modelos matemáticos: estima, predice y deduce propiedades de la población, para fenómenos que tienen cierto grado de incertidumbre.

Funciones estadísticas con R Base

Algunas funciones estadísticas

Funciones R base

Función	Descripción	Característica
mean(x)	Media	Medida de tendencia central
median(x)	Mediana	Medida de tendencia central
mode(x)	Moda	Medida de tendencia central
var(x)	Varianza	Medida de dispersión
sd(x)	Desviación estándar	Medida de dispersión
cv(x)	Coeficiente de variación	Medida de dispersión
min(x)	Mínimo	Medida de dispersión
max(x)	Máximo	Medida de dispersión
IQR(x)	Rango intercuartil	Medida de dispersión
range(x)	Rango	Medida de dispersión
quantile(x)	Cuartiles	Medida de posición

Ejemplos aplicados

```
library(tidyverse)

IMACEC <- readxl::read_excel(path = "Datos/Data_IMACEC.xlsx")
IMACEC$Year <- as.character(IMACEC$Year)

IMACEC %>% janitor::clean_names()
```

year	month	imacec	imacec_minero	imacec_no_minero	pib_minero	pib	precio_cobre
2003	12	67.76079	98.08392	63.36657	14,404.27	86,942.76	99.85
2004	1	63.70224	90.06012	59.73470	14,448.68	87,492.83	109.93
2004	2	61.32155	83.76556	57.66408	14,490.98	88,040.46	125.17
2004	3	70.11227	92.09932	66.13346	14,531.34	88,585.86	136.54
2004	4	67.97267	93.53549	63.88069	14,568.78	89,127.89	133.75
2004	5	67.89010	98.03560	63.54821	14,602.32	89,665.41	123.99
2004	6	66.47407	99.20470	62.04521	14,630.98	90,197.30	121.60
2004	7	65.67273	97.16145	61.34407	14,653.76	90,722.41	127.39
2004	8	66.65976	98.74424	62.25928	14,669.70	91,239.61	129.10

Media

```
mean(IMACEC$precio_cobre)
```

```
## [1] 282.0961
```

Mediana

```
median(IMACEC$precio_cobre)
```

```
## [1] 292.14
```

Moda

```
modeest::mlv(IMACEC$precio_cobre)
```

```
## [1] 299.4428
```

Varianza

```
var(IMACEC$precio_cobre)
```

```
## [1] 5903.072
```

Desviación estándar

```
sd(IMACEC$precio_cobre)
```

```
## [1] 76.83145
```

Coeficiente de variación

```
sd(IMACEC$precio_cobre)/mean(IMACEC$precio_cobre)
```

```
## [1] 0.2723591
```

Mínimo

```
min(IMACEC$precio_cobre)
```

```
## [1] 99.85
```

Máximo

```
max(IMACEC$precio_cobre)
```

```
## [1] 449.17
```

Rango intercuartil

```
IQR(IMACEC$precio_cobre)
```

```
## [1] 104.1825
```

Rango

```
range(IMACEC$precio_cobre)
```

```
## [1]  99.85 449.17
```

Cuartiles

```
quantile(IMACEC$precio_cobre, probs = c(0.25, 0.50, 0.75))
```

```
##       25%       50%       75%
## 233.0925 292.1400 337.2750
```

Tablas de frecuencia

Tablas de frecuencia

- `table(x)` entrega una tabla de frecuencia simple.

```
datos <- datos::paises
table(datos$continente)
```

- `prop.table(table(x))` transforma la tabla anterior en proporciones.

```
datos <- datos::paises
prop.table(table(datos$continente))
```

- `table(x, y)` permite usar más variables para tablas de dos o más entradas (tablas de contingencia).

```
datos <- datos::paises %>%
  dplyr::mutate(BajoMedia = case_when(esperanza_de_vida <= mean(esperanza_de_vida)

table(datos$continente, datos$BajoMedia)
```

Tablas de frecuencia de doble entrada

En el caso de una tabla de doble entrada, se tiene el argumento `margin` cuando se quiere representar en porcentajes. Esto permite realizar los siguientes análisis:

- `prop.table(table(x, y))`: Representa el porcentaje conjunto de obtener un elemento con la combinación deseada.
- `prop.table(table(x, y), margin = 1)`: Representa el porcentaje de obtener `y` dado que se tiene `x`.
- `prop.table(table(x, y), margin = 2)`: Representa el porcentaje de obtener `x` dado que se tiene `y`

```
prop.table(table(datos$continente, datos$BajoMedia))
```

```
##
##              Bajo Media Sobre Media
##   África    0.319248826 0.046948357
##   Américas  0.048122066 0.127934272
##   Asia      0.100938967 0.131455399
##   Europa    0.006455399 0.204812207
##   Oceanía   0.000000000 0.014084507
```

Aproximadamente, el 32% de los datos son del continente África y tienen una esperanza de vida bajo la media.

```
prop.table(table(datos$continente, datos$BajoMedia), margin = 1)
```

```
##
##              Bajo Media Sobre Media
##   África    0.87179487  0.12820513
##   Américas  0.27333333  0.72666667
##   Asia      0.43434343  0.56565657
##   Europa    0.03055556  0.96944444
##   Oceanía   0.00000000  1.00000000
```

Para los datos que provienen de África, aproximadamente el 87% de estos tienen esperanza de vida bajo la media

```r
prop.table(table(datos$continente, datos$BajoMedia), margin = 2)
```

```
##
##             Bajo Media Sobre Media
##   África    0.67243511  0.08938547
##   Américas 0.10135970  0.24357542
##   Asia      0.21260816  0.25027933
##   Europa    0.01359703  0.38994413
##   Oceanía   0.00000000  0.02681564
```

De los datos que tiene una esperanza de vida bajo la media, aproximadamente el 67% provienen de África

Análisis de correlación

Análisis de correlación

Para calcular la correlación entre dos variables numéricas, se utiliza el comando `cor(x, y)`. Además, se puede especificar el tipo de correlación con el argumento `method`.

```
flores <- datos::flores
cor(flores$Largo.Sepalo, flores$Ancho.Petalo, method = "pearson")
```

```
## [1] 0.8179411
```

Además, se puede puede ingresar una base de datos con variables numéricas, lo que retorna una matriz de correlaciones.

```
flores <- datos::flores %>%
  dplyr::select_if(is.numeric)

cor(flores, method = "pearson")
```

```
##              Largo.Sepalo Ancho.Sepalo Largo.Petalo Ancho.Petalo
## Largo.Sepalo    1.0000000   -0.1175698    0.8717538    0.8179411
## Ancho.Sepalo   -0.1175698    1.0000000   -0.4284401   -0.3661259
## Largo.Petalo    0.8717538   -0.4284401    1.0000000    0.9628654
## Ancho.Petalo    0.8179411   -0.3661259    0.9628654    1.0000000
```

En la clase anterior, se vio que una matriz de correlación puede ser representada gráficamente con la librería `ggcorrplot`.

```
ggcorrplot::ggcorrplot(cor(flores, method = "pearson"),
                       method = "circle",
                       type = "upper")
```

Resumen de datos con `dplyr`

```r
IMACEC %>%
  dplyr::filter(year >= 2015) %>%
  dplyr::group_by(year) %>%
  dplyr::summarise(media= mean(precio_cobre, na.rm = TRUE),
                   mediana = median(precio_cobre, na.rm = TRUE),
                   sd = sd(precio_cobre, na.rm = TRUE),
                   Q1 = quantile(precio_cobre, probs = 0.25),
                   Q3 = quantile(precio_cobre, probs = 0.75),
                   min = quantile(precio_cobre, na.rm = TRUE),
                   max = quantile(precio_cobre, na.rm = TRUE),
                   asimetria = moments::skewness(precio_cobre, na.rm = TRUE),
                   curtosis = moments::kurtosis(precio_cobre, na.rm = TRUE))
```

```
## # A tibble: 30 x 10
## # Groups:   year [6]
##    year  media mediana    sd    Q1    Q3   min   max asimetria curtosis
##    <chr> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>     <dbl>    <dbl>
##  1 2015   251.    255.  25.0  236.  268.  208.  208.    -0.261     2.01
##  2 2015   251.    255.  25.0  236.  268.  236.  236.    -0.261     2.01
##  3 2015   251.    255.  25.0  236.  268.  255.  255.    -0.261     2.01
##  4 2015   251.    255.  25.0  236.  268.  268.  268.    -0.261     2.01
##  5 2015   251.    255.  25.0  236.  268.  289.  289.    -0.261     2.01
##  6 2016   220.    214.  16.2  211.  221.  201.  201.     1.33      3.70
##  7 2016   220.    214.  16.2  211.  221.  211.  211.     1.33      3.70
##  8 2016   220.    214.  16.2  211.  221.  214.  214.     1.33      3.70
##  9 2016   220.    214.  16.2  211.  221.  221.  221.     1.33      3.70
## 10 2016   220.    214.  16.2  211.  221.  257.  257.     1.33      3.70
## # ... with 20 more rows
```

Taller práctico 1

Taller práctico 1

Usted está trabajando en un banco, que pretende **realizar un estudio para calcular el riesgo de crédito mediante distintos modelos estadísticos**. Sin embargo, previo a la formulación del diseño y la metodología, le solicitan realizar un **análisis exploratorio** de la siguiente base de datos:

```
data <- readxl::read_xlsx(path = "Datos/base_bancos.xlsx",
                          sheet = 1, col_names = TRUE)
head(data, 7)
```

```
## # A tibble: 7 x 11
##    Edad Sucursal Ingreso_mensual Bienes Tarjetas Credito Solicitud  Sexo
##   <dbl> <chr>              <dbl> <chr>     <dbl>   <dbl>     <dbl> <dbl>
## 1    48 C                1009411 NO            0       0         1     1
## 2    40 B                1404737 SI            1       1         0     0
## 3    51 C                 498745 SI            1       0         0     1
## 4    23 B                1309255 NO            1       0         0     1
## 5    57 A                1500255 NO            0       0         0     1
## 6    57 B                1070490 NO            1       0         1     1
## 7    22 A                1564674 NO            1       0         1     0
## # ... with 3 more variables: Profesion <chr>, Comuna <chr>, Hijos <dbl>
```

Taller práctico 1

1. Leer `base_bancos.xlsx` y realizar análisis exploratorio del conjunto de datos (naturaleza de variables, clase, nombres de columnas, resumen global, etc).

2. ¿Cuántos `clientes son profesionales`, según comuna?.

3. Agrupar `clientes bancarios` a partir de rangos etareos y ver quiénes `solicitan crédito`.

4. ¿Cuántos `hijos` tienen los clientes que `solicitan crédito`?

5. ¿Cuál es el `ingreso promedio mensual` por comuna?.

6. ¿Cuál es la `edad promedio` de hombres y mujeres que `solicitan crédito`?

Taller práctico 1

1. ¿Cuál es la mediana de `ingreso mensual` en `Lo Prado`?
2. ¿Cómo se `agrupan los ingresos` de todos los clientes?
3. ¿Entre qué `rangos` se encuentra el ingreso mensual de `San Ramón`?
4. ¿Cuál es el `rango intercuartil` de la edad de clientes en la comuna de `Providencia`?
5. ¿Hay `ingresos mensuales atípicos` en la comuna de `La Cisterna`?
6. ¿Cuál es la `forma de distribución` del ingreso?
7. ¿Cuál es el `ingreso mensual` de cada cliente, según `Comuna` (tomando como referencia forma, distribución, posición y dispersión?

Taller práctico 2

Taller práctico 2

Usted ha sido contratado como consultor/a para **ver si eventualmente el partido X inicia un estudio sobre los factores que inciden en la probabilidad de votación de X usuario de Y red social sobre una candidata a la Convención Constituyente**. En una reunión de planificación se pretende determinar preliminarmente si es que a mayor cantidad de correos de campaña electoral (a pesar de costos fijos, variables y overhead) hay mayor cantidad de conversiones (usuario realiza clicks en los enlaces para conocer a la candidata).

```
Mailings <- c(96, 40, 104, 128, 164, 76, 72, 80, 84, 180, 44, 36)
Conversiones <- c(41, 41, 51, 53, 60, 61, 50, 28, 48, 70, 33, 30)
tail(data.frame(Mailings, Conversiones), 5)
```

```
##    Mailings Conversiones
## 8        80           28
## 9        84           48
## 10      180           70
## 11       44           33
## 12       36           30
```

Referencias y material complementario

Referencias y material complementario

Aplicaciones en R

- **Link:** `R for Data Science`
- **Link:** `R in a Nutshell`
- **Link:** `Introduction to Statistical Data With R`
- **Link:** `A Handbook of Statistical Analyses Using R`
- **Link:** `Basic Statistics Using R`

¡Gracias!

Diego Muñoz Ureta
dimunoz1@uc.cl

Felipe Moya
felipe.moya@uc.cl