Análisis Estadístico I - Tabulaciones Cruzadas y Diagramas de Dispersión

Ivan Fernando Mujica Mamani   
Maestría en Ciencia de Datos  
Universidad Catolica Boliviana San PabloLa Paz, Bolivia  
ifmm87@gmail.com

***Resumen*— En el presente artículo resolveremos dos ejercicios, ambos parte del segundo capítulo del libro “Estadistica para administración y economía” con un lenguaje ampliamente usado en el campo estadístico como lo es R y gracias a Rstudio que es la herramienta de desarrollo integrado IDE por defecto de R podemos ejecutar fácilmente rutinas para el manejo de los datos en forma de arrays multidimensionales.**

***Keywords—R, Rstudio.***

# Introduccion

La Estadística se define como el arte y la ciencia de reunir datos, analizarlos, representarlos e interpretarlos. Especialmente en los negocios y en la economía [2]. La mayor parte de la información estadística en periódicos y revistas e informes de empresas, que se presentan de una forma fácil de leer, ya sea en tablas , gráficos y números se le conoce como Estadística Descriptiva[2].

R es un lenguaje ampliamente usado en el ámbito estadístico y provee una gran variedad de técnicas para el análisis estadístico y la representación gráfica de los mismos[1].

En el presente documento se muestra el proceso de tabulación y posterior graficación de una muestra con dos variables usando el lenguaje R, ambos parte de los ejercicios 29 y 30 del

# Ejercicio 29

**Los siguientes datos constan de 30 observaciones en las que intervienen dos variables, x y y. Las categorías para *x* son *A, B, C*; y para *y* son 1 y 2.**

Podemos cargar la muestra de datos con la siguiente rutina:

# Practica 1

# Leemos el archivo

install.packages("readxl")

library("readxl")

ejercicio29 <- read\_excel("/home/ivan/DATOS DEL LIBRO-ANDERSON/Ch 02 Descriptive/Crosstab.xls",'Data')

x <- ejercicio29$x # recuperamos la columna x

y <- ejercicio29$y # recuperamos la columna y

dataframe <- data.frame(x, y)

dataframe # dataframe principal

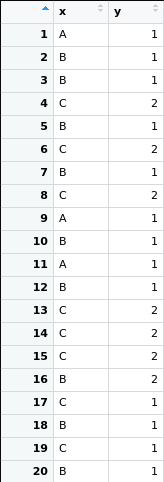


Fig. 1. Muestra de los datos

1. Con estos datos elabore una tabulación cruzada en la que *x* sea la variable para los renglones y *y* para las columnas.

Para calcular las frecuencias debemos hacer una tabulación cruzada, para ello usamos la función nativa table, el primer argumento de la función es la fila y el segundo las columnas, para obtener los totales parciales de los renglones y columnas usamos la función *addmargins* en dos ocasiones, una para los renglones y otra para las columnas

# contando las frecuencias

(sinTotales<-table(dataframe$x , dataframe$y)) # sin totales

(conTotales<-addmargins(sinTotales, margin=1)) # con totales y

(conTotales<-addmargins(conTotales, margin=2)) # con totales x

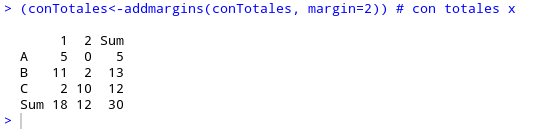


Fig. 2. Frecuencias cruzadas con totales

1. Calcule los porcentajes de los renglones

Para el cálculo de los porcentajes vamos a hacer el uso de la función prop.table, este recibe dos argumentos, el primero el *dataframe* y el segundo el eje conde se quiere aplicar la operación y adicionalmente multiplicamos por 100 para calcular el porcentaje.

b <-prop.table(sinTotales,1) # porcentaje de los renglones

(addmargins(b \* 100, margin=2))

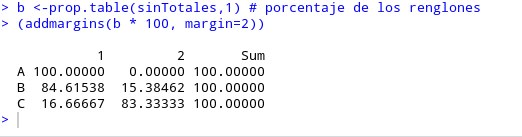


Fig. 3. Porcentaje de los renglones

1. Calcule los porcentajes de las columnas.

Al igual que para el enunciado anterior; realizamos las mismas rutinas

c <-prop.table(sinTotales,2) # porcentaje de las columnas

(addmargins(c \* 100, margin=1))

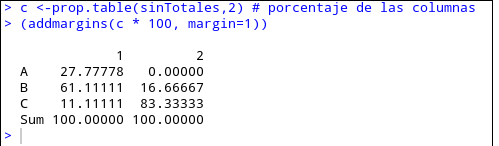


Fig. 4. Porcentaje de las columnas

1. ¿Cuál es la relación si hay alguna, entre las variables *x* y *y*?

La relación que existente entre las dos variables es la siguiente:

Los valores de la categoría *A* de *x* siempre está asociado con los valores de la categoría 1 de *Y*

Los valores categoría *B* de *x* está asociado con los valores de la categoría 1 de *y*.

Los valores de la categoría C están asociados con los valores de la categoría 2 de y.

# Ejercicio 30

**Las siguientes 20 observaciones corresponden a 20 variables cuantitativas, *x* y *y.***

install.packages("readxl")

library("readxl")

ejercicio30 <- read\_excel("/home/ivan/DATOS DEL LIBRO-ANDERSON/Ch 02 Descriptive/Scatter.xls",'Data')

x <- ejercicio30$x # recuperamos la columna x

y <- ejercicio30$y # recuperamos la columna y

dataframe30 <- data.frame(x, y)

dataframe30 # dataframe principal

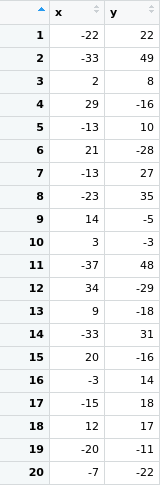


Fig. 5. Muestra de datos

1. Elabore un diagrama de dispersión

Para graficar el diagrama de dispersión usamos la función plot.

plot(x = dataframe30$x, y = dataframe30$y)

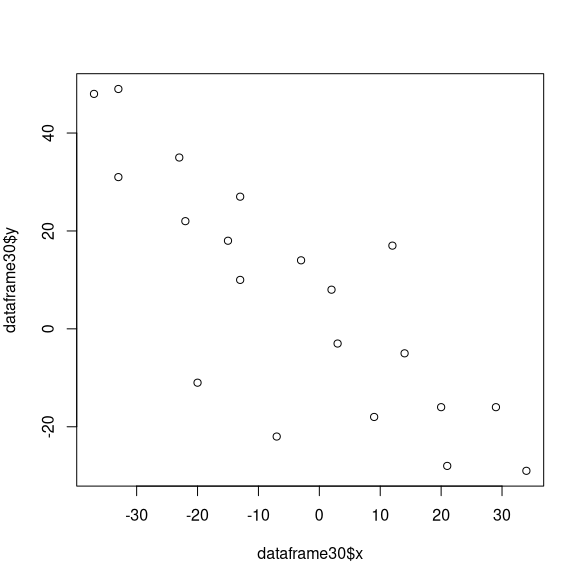


Fig. 6. Diagrama de dispersión

1. ¿Cuál es la relación, si hay alguna entre x y y?

Existe una relación negativa entre *x* y *y*, *y* se decrementa mientras x se incrementa.

##### III. **c**onclusion

En el artículo resolvimos dos ejercicios propuestos usando la función table y prop.table para la contabilización de las frecuencias cruzadas. Así también usamos la función plot para representar un diagrama de dispersión.

##### References.

1. David Anderson, Dennis Sweeney, Thomas A. Williams, Statistics for business and economy, 10th ed. Col. Santa Cruz Manca, Santa Fe: Cengage Learning Editores, pp. 2-13, 2008.
2. https://www.r-project.org/ “The R Project for Statistical Computing”[online] Available: https://www.r-project.org/ 2020.