# Análisis Estadístico de Datos Financieros con R

*Oswaldo Navarrete Carreño, María Alexandra Chávez*

*A quién aún no ha visto la luz y ya ilumina mi vida.*

# Índice general

# Capítulo 1

# ¿A quién va dirigido este libro?

Este libro no es una introducción a la estadística. En la presente obra se intenta hacer un repaso de algunos temas de estadistica que debe conocer quien desee hacer investigación en Contabilidad, en Auditoría o quizás en alguna ciencia social. Es probable que se omitan algunas cosas pero la retroalimentación de los lectores de esta obra será importante para su crecimiento.

En este texto se presentan, discuten y aplican los conceptos. La presentación de los conceptos es realizada pensando en un diálogo entre el autor y el lector, sin descuidar la formalidad de las expresiones matemáticas. Para la discusión y aplicación de los conceptos, se va mostrando al usuario como implementar el análisis estadístico en R.

Para aprovechar al máximo este libro se recomienda tener a mano una computadora con R instalado, a fin de poder ir ejecutando los códigos que se muestran. Los scripts y los conjuntos de datos que se presentan pueden ser descargados de https://github.com/oswnavarre/AEDFCR

Aunque la obra tiene un enfoque práctico, el lector no debe olvidar que aprender a usar R no implica saber estadística y que los programas estadísticos no brindan soluciones si el usuario no conoce los conceptos que deben ser aplicados.

## 1.1 Instalando R Y Rstudio

R es un lenguaje y entorno para computación estadística y gráficos. En los últimos años el uso del programa estadístico R ha ido en aumento. Puede ser descargado de https://cran.r-project.org/. Una de las características interesantes del programa es que su capacidad puede ser incrementada con la ayuda de paquetes, en la actualidad la página oficial del programa tiene cerca de 14000 paquetes.

RStudio es una interfaz que ayuda a explotar todas las capacidades de R. Rstudio se descarga de la página https://www.rstudio.com/.

# Capítulo 2

# Introducción

En nuestra vida diaria es común escuchar el término **estadística**, las tasas de desempleo, el índice de pobreza, el saldo promedio de nuestra cuenta de ahorros, el número de goles realizados en la LigaPro durante el fin de semana, etc. Aunque no es una forma incorrecta de ver las estadísticas, en este texto se pensará a la estadística como un conjunto de métodos que se utilizan para **recoger, clasificar, resumir, organizar, presentar, analizar e interpretar información numérica**

En las empresas la estadística es usada para tomar decisiones como los productos y las cantidades que deben ser producidas, la frecuencia con la que una maquinaria debe recibir mantenimiento, el tamaño del inventario, la forma de distribuir los productos, y casi todos los aspectos relativos a sus operaciones. En el estudio de las finanzas, la contabilidad, la economía y otras ciencias sociales la motivación para usar estadística radica en entender como funcionan los sistemas económicos, financieros o contables.

## 2.1 Estadística descriptiva e inferencial

El uso de la estadística puede ser de dos formas. La primera, cuando se describen y se presentan los datos. Y la segunda es cuando los datos son utilizados para hacer inferencias sobre características del ambiente o entorno de donde se seleccionaron los datos o sobre el mecanismo subyacente que generó los datos. La primera forma recibe el nombre de **estadística descriptiva** y la segunda se conoce como **estadística inferencial**

En la estadística descriptiva se utilizan métodos numéricos y gráficos para encontrar patrones y características de los datos a fin de resumir la información y presentarla de una forma significativa. Mientras que en la estadística inferencial se utilizan los daots para tomar decisiones, hacer estimaciones, pronósticos o predicciones y generalizaciones sobre el entorno del que fueron obtenidos los datos o el proceso que los generó.

Sea en estadística descriptiva o en estadística inferencial, el primer paso siempre va a ser obtener información de alguna característica, medida o valor que nos interese de un grupo de elementos. Esa característica, medida o valor de interés para el investigador recibe el nombre de **variable**.

## 2.2 Tipos de Variables

Muchos autores presentan algunas clasificaciones para las variables, sin embargo vamos a trabajar con una clasificación que se ajusta a las necesidades de la investigación en las áreas de nuestro interés. Según esta clasificación hay dos grandes grupos de variables: cuantitativas y cualitativas. Las primeras son las que toman valores **numéricos**. Mientras que las cualitativas toman valores que describen una **cualidad** o **característica**.

Las variables cuantitativas se clasifican a la vez en **continuas** que se presentan cuando las observaciones pueden tomar cualquier valor dentro de un subconjunto de los números reales, ejemplos de variables cuantitativas continuas son: edad, altura, temperatura y peso. Las **discretas** son aquellas cuya característica principal es que las observaciones pueden tomar un valor basado en un recuento de un conjunto de valores enteros distintos. Ejemplos de variables cuantitativas discretas son: número de hijos, número de comprobantes de venta emitidos en un mes, número de clientes haciendo fila durante una hora en un banco.

### 2.2.1 Niveles de medición

Hay cuatro niveles de medición **ordinal**, **nominal**, **intervalo** y de **radio**. En el nivel ordinal las observaciones toman valores que se ordenan o clasifican de forma lógica, por ejemplo las tallas de ropa (pequeña, media, grande, extra grande), la frecuencia con la que se hace una actividad (nunca, casi nunca, a veces, casi siempre, siempre). Por otro lado, en el nivel nominal las observaciones toman valores que no se pueden organizar de forna lógica, por ejemplo el sexo, el color de ojos, la marca de ropa favorita.

En el nivel de intervalo existe diferencia significativa entre valores pero el cero no representa la ausencia de la característica un ejemplo es la temperatura medida en grados Farenheit. Finalmente en el nivel de razón el 0 es significativo y la razón entre dos números es significativa, un ejemplo es la temperatura medida en grados Kelvin.

## 2.3 Primeros pasos en R

Una vez instalado R y RStudio, abrimos Rstudio para comenzar a trabajar. La ventana de RStudio se ve como se muestra en la figura 2.1.

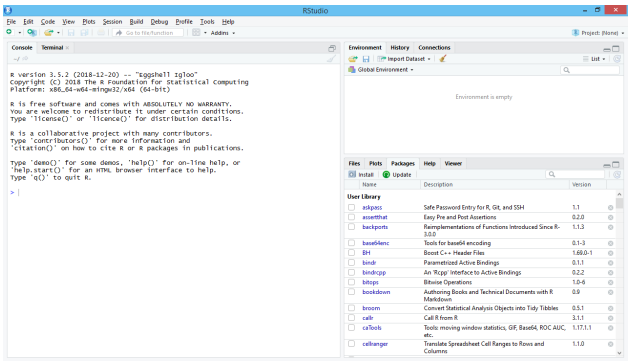


Figura 2.1: Ventana de RStudio

Lo primero que debemos hacer es abrir un nuevo script, un script de R es simplemente un archivo de texto que contiene (casi) todos los comandos que se escribirían en la línea de comandos de R, para esto en la barra de menú seguimos la secuencia **File, New File, R Script** o desde el teclado con la combinación *Ctrl + Shift + N*, en este archivo iremos escribiendo todos los comandos que vamos a trabajar. En la figura 2.2 se aprecia un script abierto.

Para empezar a aprender en el script vamos a escribir `3+2` y ejecutamos esto con la combinación de teclas **Ctrl + Enter** el resultado obviamente es 5. Ahora ingresaremos un conjunto de valores y los almacenaremos en una variable, para almacenar algo en una variable se puede usar `<-` o `=`. En la variable `x` almacenaremos un conjunto de 8 observaciones escribiendo el código:

```
x <- c(3,7,9,5,6,2,1,10)
```

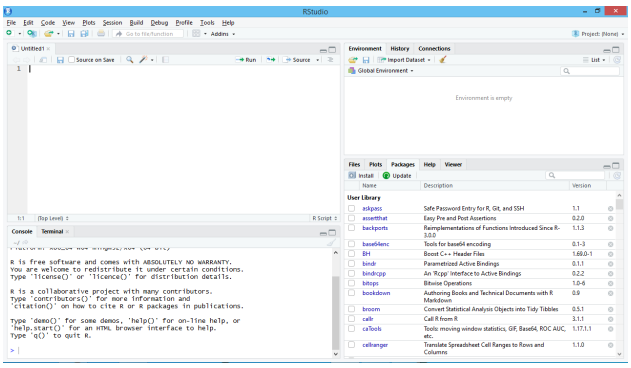

Figura 2.2: Ventana de RStudio con Script

Recuerde que este código se ejecuta con la combinación de teclas **Ctrl + Enter**. Para poder realizar análisis estadístico, es necesario cargar nuestros datos en el programa. R acepta algunos formatos de archivos, como por ejemplo archivos de Excel, archivos de valores separados por coma, archivos de texto e inclusive archivos de otros programas como SPSS. Lo más usual es trabajar con archvo de valores separados por coma es decir con extensión `.csv`, estos archivos `csv` se generan cuando el investigador recolecta la información, la almacena en un archivo de Excel o alguna otra hoja de cálculo y la guarda como un archivo de valores separados por coma.

Para trabajar de forma eficiente con R, se recomienda comenzar por fijar un directorio de trabajo donde deben estar guardados nuestros archivos en el formato que sea de nuestra preferencia. Una forma de hacerlo es desde la barra de menú **Session, Set Working Directory, Choose Directory** o desde el teclado con la combinación **Ctrl+Shift+H**, o con la función `setwd("rutadelarchivo")`.

En este primer ejercicio trabajaremos con el archivo *cap2_big4_size.csv*. Los datos serán guardados en una variable llamada `datos1`, usaremos la función `read.csv()` para leer los datos. La función `read.csv()` recibe las instrucciones `read.csv("archivo", header=T, sep=";",dec=",")`. La opción `"archivo"` indica el nombre del archivo, `header=T` o `header=F` permite indicar si las columnas tienen un encabezado que las identifique, `sep=";"` sirve para indicar cual es el separador presente en nuestro archivo en algunas ocasiones ocurre que un archivo de valores separado por coma en realidad tiene sus valores separados por un punto y cona esto generalmente ocurre cuando el sistema utiliza, como en este caso, la coma como separador decimal y finalmente la opción `dec=","` sirve para indicar que el separador decimal es la coma.

Una característica de R es que permite acceder a la ayuda sobre las funciones, esto se hace escribiendo `?funcion` por ejemplo si queremos la ayuda de la función `read.csv` simplemente escribimos `?read.csv` en el panel ubicado en la parte inferior derecha se desplegará la ayuda de la función. Con la particularidad de que la ayuda se despliega en inglés lo que no debería ser problema para un buen investigador.

El archivo que vamos a analizar contiene los activos, la utilidad, las ventas y el patrimonio de una muestra de empresas tomada de los registros de la Superintendencia de Compañías. Además en el conjunto de datos se indica si la empresa ha sido auditada por una de las 4 firmas auditoras consideradas las más grandes o también llamadas Big Four. En la 2.1 se muestran las 10 primeras observaciones de nuestro conjunto de datos.

Sin más preámbulos, empecemos a trabajar. Recapitulando, primero configuraremos el directorio de trabajo, luego cargaremos el archivo indicado. Finalmente usamos la función `str()`, la que nos permite obtener la descripción de la estructura de los datos.

```
setwd("C:/Users/onava_000/OneDrive/libro_mc/estadistica")
datos1 <- read.csv("cap2_big4_size.csv",header=TRUE,sep=";",dec=",")
str(datos1)
```

```
## 'data.frame':    2256 obs. of  6 variables:
```

Tabla 2.1: Primeras 10 observaciones

| EXPMUESTRA | BIG4 | ACTIVOS | UTILIDAD | VTAS | PAT |
|---|---|---|---|---|---|
| 85 | 1 | 73315618 | 7522758.7 | 191474544 | 39382529 |
| 100121 | 0 | 21052702 | -122898.5 | 132585022 | 1577764 |
| 45178 | 0 | 10468672 | 536876.9 | 13974269 | 4312094 |
| 51193 | 0 | 4130483 | 455759.4 | 8670153 | 1858990 |
| 47598 | 0 | 23507401 | 266370.5 | 18555609 | 7137609 |
| 31720 | 0 | 7220312 | 437718.3 | 16097135 | 4002154 |
| 46189 | 0 | 14526822 | 1206400.9 | 12281188 | 5015806 |
| 9731 | 0 | 8539445 | 367848.1 | 10844918 | 2232339 |
| 4619 | 0 | 2605059 | -22438.4 | 6244589 | 1366610 |
| 102434 | 0 | 23975816 | 790265.6 | 40612649 | 8754369 |

```
##  $ EXPMUESTRA: int  85 100121 45178 51193 47598 31720 46189 9731 4619 102434 ...
##  $ BIG4      : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ ACTIVOS   : num  73315618 21052702 10468672 4130483 23507401 ...
##  $ UTILIDAD  : num  7522759 -122898 536877 455759 266371 ...
##  $ VTAS      : num  1.91e+08 1.33e+08 1.40e+07 8.67e+06 1.86e+07 ...
##  $ PAT       : num  39382529 1577764 4312094 1858990 7137609 ...
```

En la primera línea de los resultados se observa la salida `'data.frame': 2256 obs. of 6 variables:` esto nos indica que nuestro *marco de datos (data frame)* tiene 2256 observaciones y 6 variables. Con respecto a las variables tenemos 6 variables que a continuación se describen y se explican los resultados obtenidos con la función.

- `EXPMUESTRA`: esta variable es de tipo entera (INT) y almacena el expediente de la empresa. Aunque la variable tiene valores numéricos, no es una variable cuantitativa sino cualitativa "Expediente de la Empresa"
- `BIG4`: esta variable es de tipo entera, y ha sido codificada con 1 si la empresa fue auditada por una Big Four y 0 si no. Podemos cambiar esta codificación por "Sí" y "No" en lugar de "1" y "0", más adelante aprendemos como hacerlo. Al igual que la variable anterior aunque tiene valores numéricos, no es una variable cuantitativa sino cualitativa, dejamos al lector la reflexión en este particular.
- `ACTIVOS`: contiene el valor de los activos totales de la empresa. Es de tipo `num` porque permite el uso de decimales. Corresponde a una variable cuantitativa continua.
- `UTILIDAD`: contiene el valor de la utildad de la empresa.
- `VTAS`: contiene el valor de las ventas de la empresa.
- `PAT`: contiene el valor del patrimonio de la empresa.

Los paquetes de R son colecciones de funciones y conjuntos de datos desarrollados por la comunidad de usuarios, los paquetes aumentan el poder de R mejorando las funcionalidades existentes en la base de R, o añadiendo nuevas funcionalidades. En este texto trabajaremos con algunos de los paquetes desarrollados por el equipo de RStudio, una descripción detallada de estos paquetes puede ser encontrada en https://www.rstudio.com/products/rpackages/. . Comenzaremos por instalar el paquete `dplyr`, este paquete tiene funciones que permiten realizar facilmente manipulaciones de datos. Para instalar un paquete se utiliza la función `install.packages("paquete")`. Una vez instalado el paquete, se carga el paquete utilizando la función `library(paquete)`.

```
install.packages("dplyr")
```

La primera manipulación que vamos a realizar es la creación de nuevas variables con el paquete `dplyr`. En nuestros datos cargados en el conjunto de datos `datos1` vamos a crear tres variables nuevas **ROA**, **ROS** y **ROE**. Recordemos que el **Retorno sobre activos** ( **ROA**, Return on Assets) se lo calcula como la razón entre la utilidad y los activos como se ve en la ecuación (2.1). En las ecuaciones (2.2) y (2.3) se dan las expresiones para calcular el **Retorno sobre ventas** ( **ROS** Return on Sales) y el **Retorno sobre el**

**Patrimonio** ( **ROE** Return on Equity)

$$ROA = \frac{Utilidad}{Activos} \tag{2.1}$$

$$ROS = \frac{Utilidad}{Ventas} \tag{2.2}$$

$$ROE = \frac{Utilidad}{Patrimonio} \tag{2.3}$$

Una característica importante de `dplyr` es el uso del operador `%>%`. Cada transformación u operación en los datos se separa por el operador `%>%`. La primera función de `dplyr` que usaremos es `mutate()`, básicamente esta función permite crear nuevas variables.

```
library(dplyr)
datos1 <- datos1 %>%
  mutate(
    ROA = UTILIDAD/ACTIVOS,
    ROS = UTILIDAD/VTAS,
    ROE = UTILIDAD/PAT
  )
str(datos1)
```

```
## 'data.frame':    2256 obs. of  9 variables:
##  $ EXPMUESTRA: int  85 100121 45178 51193 47598 31720 46189 9731 4619 102434 ...
##  $ BIG4      : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ ACTIVOS   : num  73315618 21052702 10468672 4130483 23507401 ...
##  $ UTILIDAD  : num  7522759 -122898 536877 455759 266371 ...
##  $ VTAS      : num  1.91e+08 1.33e+08 1.40e+07 8.67e+06 1.86e+07 ...
##  $ PAT       : num  39382529 1577764 4312094 1858990 7137609 ...
##  $ ROA       : num  0.10261 -0.00584 0.05128 0.11034 0.01133 ...
##  $ ROS       : num  0.039289 -0.000927 0.038419 0.052566 0.014355 ...
##  $ ROE       : num  0.191 -0.0779 0.1245 0.2452 0.0373 ...
```

En las últimas líneas de la salida de R, se observa que ahora en el conjunto de datos existen ahora tres nuevas variables. En la próxima sección seguiremos trabajando con el mismo conjunto de datos.

## 2.4 Medidas de Tendencia Central

Una medida de tendencia central, es una medida de resumen que intenta describir un conjunto completo de datos con un único valor que representa la mitad o centro de la distribución.

Las tres medidas de tendencia central principales son la media la mediana y la moda.

### 2.4.1 Media

La media se la calcula como la suma de todos los valores de una variable dividido para el número de valores. En la ecuación (2.4) se muestra la fórmula para calcular la media.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.4}$$

La media tiene algunas propiedades que a continuación se detallan:

- Si a cada valor $x_i$ de una distribución con media $\bar{x}$ se le suma un valor constante $k \in \mathbb{R}$, la nueva media es $\bar{x} + k$
- Si a cada valor $x_i$ de una distribución con media $\bar{x}$ se lo multiplica por un valor constante $k \in \mathbb{R}$, la nueva media es $k\bar{x}$
- Si a cada valor $x_i$ de una distribución con media $\bar{x}$ se lo divide por un valor constante $k \neq 0 \in \mathbb{R}$, la nueva media es $\dfrac{\bar{x}}{k}$

Las ventajas de usar la media son:

- Es fácil de entender y calcular
- No se ve afectada mayormente por fluctuaciones productos del muestreo
- Toma en cuenta todos los valores de la variable

Las desventajas de usar la media son:

- Es muy sensible a la presencia de pocos valores muy pequeños o muy grandes, dicho de otra forma la media es sensible a valores aberrantes.
- No se puede calcular por inspección.

### 2.4.2 Mediana

La mediana es el valor central en una distribución cuando se ordenan los valores de forma ascendente o descendente. El valor de la mediana depende entonces del número de valores presentes en la variable. Definamos como $\{X\}$ al conjunto de datos ordenado, y sea $\{X\}_i$ el valor i-ésimo del conjunto $\{X\}$ entonces la mediana $Me$ se define como

$$Me = \begin{cases} \{X\}_{\frac{n+1}{2}} & ; n \quad \text{impar} \\ \dfrac{\{X\}_{\frac{n}{2}} + \{X\}_{\frac{n}{2}+1}}{2} & ; n \quad \text{par} \end{cases} \tag{2.5}$$

Lo escrito en la ecuación (2.5) se puede expresar de la siguiente forma: si el número de datos es impar, la mediana es igual al valor central de la distribución y si el número de datos es par, la mediana es igual al promedio de los valores centrales de la distribución.

Las ventajas de usar la mediana son:

- Es fácil de calcular y comprender
- No se ve afectada por valores extremos
- Se puede determinar para escalas ordinales, nominales, de razón e intervalo

Las desventajas de usar la mediana son:

- No toma en cuenta el valor exacto de cada dato y por tanto no usa toda la información disponible.
- Si se agrupan los valores de dos grupos, la mediana de cada grupo no puede ser expresada en términos del grupo agrupado.

### 2.4.3 Moda

La moda es definida como el valor que ocurre con mayor frecuencia en los datos. Algunos conjuntos de datos no tienen moda porque cada valor ocurre solo una vez. Hay conjuntos de datos que tienen más de una moda, si tienen 2 modas reciben el nombre de bimodal y se acostumbra que si tiene más de 3 modas se la llama multimodal.

Las ventajas de usar la moda son:

- Puede ser usada para datos con escala nominal

- Es sencilla de calcular

La desventaja de la moda es:

- No es usada en análisis estadístico debido a que no está definida algebraicamente y la fluctuación en la frecuencia de las observaciones es mayor cuando el tamaño de la muestra es pequeña.

### 2.4.4 ¿trabajamos con la media o la mediana?

La media es considerada generalmente la mejor medida de tendencia central y la más usada. Sin embargo, hay situaciones donde las otras medidas de tendencia central son preferidas.

La mediana es preferida a la media cuando:

- Hay valores extremos en la distribución
- Hay valores indeterminados
- Los datos son medidos en una escala ordinal

La moda es la medida preferida cuando los datos son medidos en una escala nominal.

### 2.4.5 Cálculo de las medidas de tendencia central en R

Para calcular la media y la mediana se utilizan las funciones `mean()` y `median()` respectivamente, estas dos funciones vienen cargadas con los paquetes base de R. Para calcular la moda usaremos la función `Mode()` del paquete `DescTools`, recuerde que para instalar un paquete se utiliza la función `install.packages()`.

En el siguiente ejemplo se obtiene la media de los activos de las empresas. como solamente necesitamos una variable del conjunto de datos usamos el operador `$`, el funcionamiento de este operador es `data.frame$variable` es decir indicamos el conjunto de datos del que llamamos la variable y después del operador `$` indicamos la variable que vamos a trabajar.

```
mean(datos1$ACTIVOS)
```

```
## [1] 44064165
```

```
median(datos1$ACTIVOS)
```

```
## [1] 10326361
```

```
library(DescTools)
Mode(datos1$ACTIVOS)
```

```
## [1]  55996406 628446149
```

En el resultado de la moda se obtienen 2 valores. Es decir que existen dos valores que se repiten más veces o tienen mayor frecuencia. Cuando se realiza investigación es común desear hacer una tabla con las estadísticas descriptivas de los datos. El paquete `dplyr` permite realizar tablas que resuman las variables de forma sencilla con la función `summarise()`.

```
datos1 %>%
  summarise(PROM.ACTIVOS = mean(ACTIVOS),
            PROM.UTILIDAD = mean(UTILIDAD),
            PROM.VTAS = mean(VTAS),
            MEDIAN.ACTIVOS = median(ACTIVOS),
            MEDIAN.UTILIDAD = median(UTILIDAD),
            MEDIAN.VTAS = median(VTAS)
            )
```

```
##   PROM.ACTIVOS PROM.UTILIDAD PROM.VTAS MEDIAN.ACTIVOS MEDIAN.UTILIDAD
## 1     44064165       4250664  50555030       10326361        350642.1
##   MEDIAN.VTAS
## 1     9190661
```

## 2.5 Tablas de frecuencia

Una tabla de frecuencia es una forma de describir los datos

## 2.6 Tablas de Contingencia

## 2.7 Gráficos y Visualización

### 2.7.1 Diagramas de Caja y valores atípicos

#### 2.7.1.1 Intervalos de Confianza

## 2.8 Medidas de dispersión

# Capítulo 3

# Pruebas de Hipótesis

# Capítulo 4

# Regresión

# Capítulo 5

# Análisis Factorial

## 5.1 Análisis de Fiabilidad

## 5.2 Evaluación de Análisis Factorial

# Capítulo 6

# Algo de series de tiempo

# Bibliografía

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.9.