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.

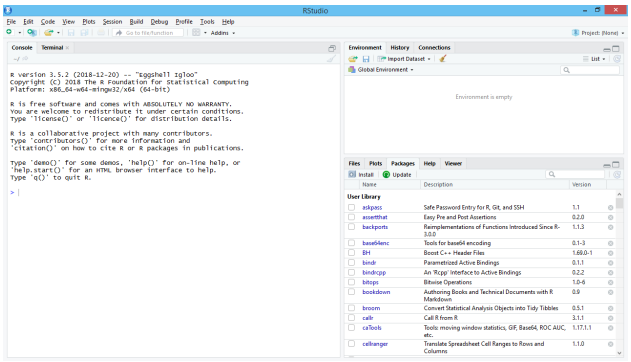

Figura 2.1: Ventana de RStudio

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)
```

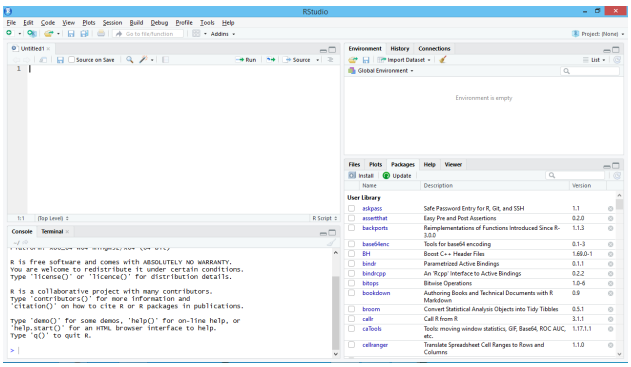

Figura 2.2: Ventana de RStudio con Script

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 2.3 se muestran las 10 primeras observaciones de nuestro conjunto de datos.

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

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")
big4size <- read.csv("cap2_big4_size.csv",header=TRUE,sep=";",dec=",")
str(big4size)
```

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

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)
big4size <- big4size %>%
  mutate(
    ROA = UTILIDAD/ACTIVOS,
    ROS = UTILIDAD/VTAS,
    ROE = UTILIDAD/PAT
```

```
  )
str(big4size)
```

```
## '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(big4size$ACTIVOS)
```

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

```
median(big4size$ACTIVOS)
```

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

```
library(DescTools)
Mode(big4size$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()`.

```
big4size %>%
  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 Medidas de posición (Cuantiles)

Las medidas de posición no central permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Entre las medidas de posición no central más importantes están los cuantiles. El término cuantil fue usado por primera vez por Kendall en 1940.

El cuantil de orden p de una distribución con $0 < p < 1$ es el valor x_i de la variable X que marca un corte de modo que una proporción p o un porcentaje $100p\%$ de valores de la población es menor o igual que x_i Por ejemplo el cuantil de orden 0.35 dejaría un 35% de valores por debajo de él.

2.5.1 Tipos de Cuantiles

- *Cuartiles*: son 3 valores (Q_1, Q_2, Q_3) que dividen a la distribución en 4 partes iguales.

- *Quintiles*: son 4 valores (K_1, K_2, K_3, K_4) que dividen a la distribución en 5 partes iguales.
- *Deciles*: son 9 valores $(D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9)$ que dividen a la distribución en 10 partes iguales.
- *Percentiles*, son 99 valores $(P_1, P_2, \ldots P_{99})$ que dividen a la distribución en 100 partes iguales.

2.5.2 Cálculo de cuantiles

Es fácil darse cuenta que existen equivalencias importantes entre los cuantiles, algunos ejemplos de estas equivalencias:

- $D_5 = Q_2 = P_{50}$
- $D_4 = K_2 = P_{40}$
- $D_3 = P_{30}$

Se deduce entonces que no es necesario tener una expresión para cada tipo de cuantiles, basta con conocer una expresión para calcular percentiles. Para esto debemos conocer dos cosas:

1. La posición del percentil en nuestro conjunto de datos.
2. El valor del percentil tomando en cuenta su posición.

Para calcular la posición del percentil i que acumula el $100p\%$ en un conjunto de datos no agrupado X, de tamaño n y ordenado en forma ascendente primero determinamos la posición del percentil con la expresión:

$$Posicin = p(n-1) + 1 \tag{2.6}$$

Para determinar el valor $X_{i.a}$ utilizamos la expresión:

$$X_{i.a} = X_i + 0.a(X_{i+1} - X_i) \tag{2.7}$$

Para calcular percentiles en R, se utiliza la función `quantile()`. Esta función recibe dos argumentos, la variable de la que se calcula el percentil y el porcentaje del percentil que se desea calcular. Se pueden calcular varios percentiles al mismo tiempo.

Vamos a calcular el primer cuartil Q_1 de la variable `ACTIVOS` del conjunto de datos ya trabajado anteriormente. Vamos a llamar a esta variable utilizando la notación `$` esta notación se usa poniendo `data.frame$variable` en este caso nuestra variable está en el conjunto `datos1`y se llama `ACTIVOS` por lo que para llamar la variable desde la función escribimos `datos1$ACTIVOS`. Luego debemos recordar que $Q_1 = P_{25}$ es decir que en la función `quantile` debemos anotar 0.25

```
quantile(big4size$ACTIVOS, 0.25)
```

```
##     25%
## 3184669
```

Ahora calculamos los tres cuartiles en este caso podemos escribir dentro de una lista los tres valores, para ingresar listas en R lo hacemos con `c(elemento1, elemento2, ... )`

```
quantile(big4size$ACTIVOS, c(0.25,0.50,0.75))
```

```
##      25%      50%      75%
##  3184669 10326361 33192848
```

De los resultados obtenidos se interpreta que el 25% de los activos de las empresas es menor que 3184669. Supongamos que se quieren determinar los deciles, una forma de hacer la lista es con la función `seq` con las instrucciones `seq(inicial, final, by = aumento)` de esta manera evitamos escribir los nueve valores.

```
quantile(big4size$ACTIVOS, seq(0.1,0.9, by = 0.1))
```

```
##       10%       20%       30%       40%       50%       60%       70%       80%
##   1621865   2561491   3882643   6187167  10326361  16883801  26613778  49838668
##       90%
## 93545755
```

2.6 Medidas de dispersión

Si comparamos los conjuntos de datos $X = \{2, 4, 6, 8\}$ y $Y = \{1, 3, 7, 9\}$ se obtiene que las medias son iguels $\bar{X} = \bar{Y} = 5$. En la figura 2.3 se ha graficado con color naranja los puntos del conjunto X y de color azul los puntos del conjunto Y. Se observa que los valores del conjunto Y están más dispersos que los valores del conjunto X. En esta sección se discute las formas existentes para cuantificar la dispersión.

Figura 2.3: Conjuntos graficados

2.6.1 Rango

El rango es la medida de dispersión más fácil de calcular. Se obtiene restando el máximo menos el mínimo. La expresión para calcularlo es:

$$Rango = max - min \tag{2.8}$$

2.6.2 Varianza

La varianza es el promedio de la diferencia de la media cuadrática. Si se conocen todos los datos de una población se puede calcular la varianza poblacional:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x - \mu)^2}{N} \tag{2.9}$$

Por otro lado si se conocen los datos de una población se puede calcular la varianza muestral:

$$s^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n - 1} \tag{2.10}$$

2.6.3 Desviación

La desviación es la raíz cuadrada de la varianza, en las fórmulas (2.11) y (2.12) se muestran las expresiones para calcular la desviación poblacional y muestral respectivamente.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x - \mu)^2}{N}} \tag{2.11}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n-1}} \tag{2.12}$$

2.6.4 Medidas de dispersión en R

Es necesario saber que R defecto no tiene una función para calcular el rango sin embargo para calcular el rango vamos a usar `max() - min()`, y que además por defecto R tiene una función para calcular la varianza muestral (`var()`) y otra para calcular la desviación muestral `sd()`, si se desea obtener la varianza y la desviación poblacional existen por lo menos 3 soluciones:

- Se puede multiplicar la varianza muestral por $\frac{n-1}{n}$ para obtener la varianza poblacional y la desviación muestral por $\sqrt{\frac{n-1}{n}}$ para obtener la desviación poblacional.
- Se puede multiplicar la varianza muestral por $\frac{n-1}{n}$ para obtener la varianza poblacional y a ese resultado extraer la raíz cuadrada para obtener la desviación poblacional.
- Crear funciones propias que calculen la varianza y la desviación muestral.

Vamos a trabajar con la segunda solución que es simplemente una mejora de la primera solución, la tercera solución es avanzada y será abordada más adelante.

A manera de ejemplo vamos a calcular las medidas de dispersión de los activos en millones de dólares de la base `cap2_big4_size.csv`. Se calculan la varianza y la desviación poblacional aunque, a menos de que tengamos todos los datos (población), siempre en el análisis estadístico de datos se calcula la varianza y la desviación muestral.

```
big4size %>%
  summarise(RANGO.ACTIVOS = max(ACTIVOS/1000000)-min(ACTIVOS/1000000),
            VARM.ACTIVOS = var(ACTIVOS/1000000),
            DESVM.ACTIVOS = sd(ACTIVOS/1000000),
            n=n()
            ) %>%
  mutate(VARP.ACTIVOS = VARM.ACTIVOS*((n-1)/n),
         DESVP.ACTIVOS = sqrt(VARP.ACTIVOS)) %>%
  select(RANGO.ACTIVOS, VARM.ACTIVOS, DESVM.ACTIVOS, VARP.ACTIVOS, DESVP.ACTIVOS)
```

```
##   RANGO.ACTIVOS VARM.ACTIVOS DESVM.ACTIVOS VARP.ACTIVOS DESVP.ACTIVOS
## 1      1341.989     11327.34        106.43     11322.31      106.4064
```

2.7 Tablas de frecuencia

Una tabla de frecuencia es una forma de describir los datos de forma resumida, las tablas de frecuencia pueden construirse para variables cualitativas y para variables cuantitativas.

2.7.1 Variables Cualitativas

Para las variables cualitativas una tabla de frecuencia basicámente tiene tres columnas: “Categoría”, “Frecuencia”, “Porcentaje”. Para aprender a realizar tablas de frecuencia para variables cualitativas, trabajaremos con el conjunto de datos `audit_bolsa`, Este conjunto de datos tiene información sobre las empresas que cotizan en la Bolsa de Valores de Guayaquil, se elaborará una tabla de frecuencias de las firmas auditoras que han trabajado para estas empresas. La variable en la que se almacena esta información es la variable `FIRMA`.

La tabla de frecuencia se elabora usando el paquete `dplyr`. El comando `mutate( )` sirve para crear nuevas columnas, en este caso se crea la columna porcentaje.

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

tabla_firma <- audit_bolsa %>%
  group_by(FIRMA) %>%
  summarise(Frecuencia=n()) %>%
  mutate(Porcentaje = round(100*Frecuencia/sum(Frecuencia),2)
         ) %>%
  arrange(desc(Porcentaje))
 print(tabla_firma)
```

```
## # A tibble: 15 x 3
##    FIRMA                                Frecuencia Porcentaje
##    <fct>                                     <int>      <dbl>
##  1 DELOITTE                                    103      48.6
##  2 MOORE STEPHENS                               29      13.7
##  3 PWC PRICE WATER HOUSE COOPERS                24      11.3
##  4 HANSEN HOLM & CO. CIA. LTDA.                 15       7.08
##  5 KPMG                                         13       6.13
##  6 ERNST & YOUNG                                 7       3.3
##  7 BDO                                           6       2.83
##  8 ALTAMIRANO HIDALGO MARIO ROBERTO              3       1.42
##  9 KRESTON                                       3       1.42
## 10 PKF                                           3       1.42
## 11 BATALLAS & BATALLAS                           2       0.94
## 12 ASE + ASESORANDO MAS                          1       0.47
## 13 CONSULTORES MORAN CEDILLO CIA. LTDA           1       0.47
## 14 HERRERA CHANG 6 ASOCIADOS                     1       0.47
## 15 NGV                                           1       0.47
```

Tabla de Frecuencia de Firmas Auditoras

FIRMA

Frecuencia

Porcentaje

DELOITTE

103

48.58

MOORE STEPHENS

29

13.68

PWC PRICE WATER HOUSE COOPERS

24

11.32

HANSEN HOLM & CO. CIA. LTDA.

15

7.08

KPMG

13

6.13

ERNST & YOUNG

7

3.30

BDO

6

2.83

ALTAMIRANO HIDALGO MARIO ROBERTO

3

1.42

KRESTON

3

1.42

PKF

3

1.42

BATALLAS & BATALLAS

2

0.94

ASE + ASESORANDO MAS

1

0.47

CONSULTORES MORAN CEDILLO CIA. LTDA

1

0.47

HERRERA CHANG 6 ASOCIADOS

1

0.47

NGV

1

0.47

En la tabla 2.7.1 se aprecia el resultado obtenido y formateado para ser publicado. El resultado de R, puede ser exportado a un archivo Excel con la finalidad de luego tomar esa tabla y llevarla a un documento donde se

presentará toda la información analizada. Para exportar la información a un archivo excel se puede trabajar con el paquete `xlsx`. Para exportar los resultados a Excel se puede proceder de la siguiente forma.

1. Cargar el paquete `xlsx`.
2. Convertir el resultado a un `data frame` utilizando la función `as.data.frame()`
3. Exportar el resultado con la función `write.xlsx()` cuya estructura básica es `write.xlsx(datos, "archivo.xlsx")`, si se desea consultar más detalles de la función se puede escribir `?write.xlsx`.

El resultado de esta operación será un archivo de excel guardado en nuestro directorio de trabajo.

```
library(xlsx)
tabla_firma = as.data.frame(tabla_firma)
write.xlsx(tabla_firma, "tablas.xlsx", sheetName = "firmas", row.names = FALSE)
```

La opción `sheetname = "firmas"` crea dentro del libro `tablas.xlsx` una hoja de cálculo llamada `firmas`. La opción `row.names = FALSE` hace que en el archivo final no se graben los números de cada fila.

Nota: es importante tener fijado el directorio de trabajo, como se explicó en la sección 2.3.

2.7.2 Variables Cuantitativas

Una tabla de frecuencias para variables cualitativas tiene 6 columnas:

1. Clase: una clase es un intervalo del tipo $[menor, mayor)$
2. Marca de Clase: es un valor igual al promedio de los dos extremos de la clase.
3. Frecuencia: la frecuencia es igual al número de valores de la variable que están dentro del intervalo.
4. Frecuencia relativa: la frecuencia relativa se la calcula como la frecuencia dividida para el total de valores de la variable.
5. Frecuencia acumulada: se la calcula sumando las frecuencias desde la primera clase hasta la clase en consideración.
6. Frecuencia Relativa acumulada: se la calcula como la frecuencia acumulada pero para las frecuencias relativas.

Una de las ventajas de usar R es que se pueden crear funciones para cada necesidad que el investigador tenga, en este caso el código que se muestra sirve para hacer tablas de frecuencia de cualquier variable cuantitativa. A manera de ejemplo se hará la tabla de frecuencia de la variable `VTAS` en millones de dólares, del conjunto de datos trabajado en la sección 2.3.

```
library(agricolae)
library(dplyr)

h2<-with(big4size,graph.freq(VTAS/1000000,plot=FALSE));

h2 = table.freq(h2)

h3 <- h2 %>%
  mutate(Clase = paste("[",Lower,",",Upper,")"),
         "Marca de Clase"  =  Main,
         Frec. = Frequency,
         "Frec. Rel." = Percentage,
         "Frec. Acu." = CF,
         "Rel. Acu." = CPF )  %>%
  select(-c(1:7))
```

Tabla de Frecuencia de las Ventas

Clase

Marca de Clase

Frec.

Frec. Rel.

Frec. Acu.

Rel. Acu.

[0 , 165.76)

82.88

2099

93.0

2099

93.0

[165.76 , 331.52)

248.64

83

3.7

2182

96.7

[331.52 , 497.28)

414.40

35

1.6

2217

98.3

[497.28 , 663.04)

580.16

17

0.8

2234

99.0

[663.04 , 828.8)

745.92

0

0.0

2234

99.0

[828.8 , 994.56)

911.68

7

0.3

2241

99.3

[994.56 , 1160.32)

1077.44

12

0.5

2253

99.9

[1160.32 , 1326.08)

1243.20

0

0.0

2253

99.9

[1326.08 , 1491.84)

1408.96

0

0.0

2253

99.9

[1491.84 , 1657.6)

1574.72

0

0.0

2253

99.9

[1657.6 , 1823.36)

1740.48

2

0.1

2255

100.0

[1823.36 , 1989.12)

1906.24

1

0.0

2256

100.0

De la tabla 2.7.2 se obsserva que el 93% de las empresas realiza ventas entre 0 y 165.76 millones. El 99% de las empresas es decir 2234 tiene ventas menores a 663.04 millones de dólares, esto se lo puede ver en la columna de frecuencias acumuladas relativas. Además, solo una empresa tiene ventas entre 1823.36 y 1989.12 millones. Finalmente vamos a exportar la tabla de frecuencia en el archivo `tablas.xlsx`. La opción `append = TRUE` sirve para añadir una nueva hoja de cálculo al libro.

```
library(xlsx)
h3 = as.data.frame(h3)
write.xlsx(h3, "tablas.xlsx", sheetName = "frec_ventas", row.names = FALSE,append=TRUE)
```

2.8 Tablas de Contingencia

Una tabla de contingencia es una forma útil para examinar relaciones entre dos variables categóricas. Los valores en las celdas de una tabla de contingencia pueden ser de frecuencia absoluta o frecuencia relativa.

Para ejemplificar la construcción de una tabla de contingencia vamos a trabajar con el archivo `Ranking2018Guayas.csv` este archivo contiene información sobre una muestra de 162 empresas de la provincia del Guayas. Se analizará la relación entre la ciudad y el tamaño de las empresas.

```
setwd("C:/Users/onava_000/OneDrive/libro_mc/estadistica")
rank2018 = read.csv("Ranking2018Guayas.csv",header=TRUE, sep=";")


ciudad.tama = rank2018 %>%
  group_by(CIUDAD, TAMAÑO)%>%
  summarise(n=n())%>%
  spread(TAMAÑO, n) %>%
  replace(., is.na(.), 0)
```

Tabla de Contingencia de las empresas clasificadas por tamaño y ciudad

CIUDAD

MEDIANA

MICROEMPRESA

PEQUEÑA

DAULE

1

1

0

ELOY ALFARO

0

2

0

GUAYAQUIL

2

117

28

MILAGRO

0

2

0

NARANJITO

0

4

0

SAMBORONDÓN

0

0

1

SANTA LUCIA

0

1

0

VELASCO IBARRA

0

1

1

En la tabla 2.8 se observa que de las 162 empresas 117 son microempresas y de la ciudad de Guayaquil, de la ciudad de Samborondón se ha tomada una empresa pequeña. Esta información, como se mencionó antes, puede también ser mostrada en porcentajes. En la tabla 2.8 se observa la tabla de contingencia con los porcentajes.

```
ciudad.tama.porc = rank2018 %>%
  group_by(CIUDAD, TAMAÑO)%>%
  summarise(Porc = round(100*n()/nrow(rank2018),2)) %>%
  spread(TAMAÑO, Porc) %>%
  replace(., is.na(.), 0)
```

Tabla de Contingencia de las empresas clasificadas por tamaño y ciudad

CIUDAD

MEDIANA

MICROEMPRESA

PEQUEÑA

DAULE

0.62

0.62

0.00

ELOY ALFARO

0.00

1.24

0.00

GUAYAQUIL

1.24

72.67

17.39

MILAGRO

0.00

1.24

0.00

NARANJITO

0.00

2.48

0.00

SAMBORONDÓN

0.00

0.00

0.62

SANTA LUCIA

0.00

0.62

0.00

VELASCO IBARRA

0.00

0.62

0.62

2.9 Gráficos y Visualización

Para realizar gráficas R tiene algunos paquetes disponibles, sin embargo en este texto trabajaremos con el paquete `ggplot2`. Este paquete está basada en la gramática de los gráficos (Wilkinson, 2005).

2.9.1 Histogramas

Los histogramas se utilizan para variables continuas. Un histograma es un gráfico de la distribución de frecuencia de una variable, en el eje vertical se representa la frecuencia (absoluta o relativa) y en el eje horizontal los rangos de los valores.

En la figura 2.4 se muestra el histograma de la variable ventas en millones de dólares del archivo *cap2_big4_size.csv* ya descrito en la sección 2.3, este primer histograma ha sido configurado para presentar 12 barras, que las barras sean de color azul con un contorno rojo. Antes de abordar los detalles mencionados discutiremos brevemente el funcionamiento de la gramática de `ggplot2`, una gráfica realizada en `ggplot2` empieza por `ggplot(data, aes())` dentro de `aes()` se indica las variables que van a intervenir en la gráfica, Luego se añade la `geom` con la que se va a trabajar en este caso se escogió `geom_histogram()` puesto que se desea realizar un histograma. Como se indicó anteriormente se configuró el histograma con 12 barras (`bins=12`), la opción `color="red"` permite que el contorno de las barrras sea rojo y la opción `fill="blue"`hace que las barras sean de color azul.

```
library(ggplot2)

ggplot(big4size, aes(x= VTAS/1000000)) +
  geom_histogram(bins=12, color= "red", fill="blue" )
```

Para configurar las etiquetas de los ejes podemos añadir las opciones `xlab( )` y `ylab( )`. En la figura 2.5 se aprecia el histograma con las etiquetas de los ejes añadidos.

```
library(ggplot2)

ggplot(big4size, aes(x= VTAS/1000000)) +
  geom_histogram(bins=12, color= "red",  fill="blue" ) +
  xlab("Ventas en Millones de Dólares") + ylab("Frecuencia")
```

Usando el archivo `Ranking2018Guayas.csv`, vamos ahora a hacer el histograma de las ventas en miles de acuerdo al tamaño de la empresa. En la figura 2.6 se observa el histograma.

Para elaborar este histograma se tomaron en cuenta varias cosas, lo primero se estimaron los valores máximo y mínimo de la variable.

```
min(rank2018$VENTAS/1000)
## [1] 0
max(rank2018$VENTAS/1000)
## [1] 1347.729
```

Los valores obtenidos para el máximo y el mínimo fueron 0 y 1348 respectivamente, por esta razón se decidió crear 10 clases y cada clase con una longitud de 150. Adicionalmente para obtener un gráfico agradable a la vista se cambia la orientación de las marcas de 0° a 90° en el eje x con la instrucción `theme(axis.text.x = element_text(angle = 90, hjust = 1))`.

```
ggplot(rank2018, aes(x=VENTAS/1000, fill=TAMAÑO)) +
  geom_histogram(alpha=0.3, color="black",bins=10, binwidth = 150) +
  scale_x_continuous(breaks = seq(0,1350,150)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  xlab("Ventas en Miles") + ylab("Frecuencia")
```

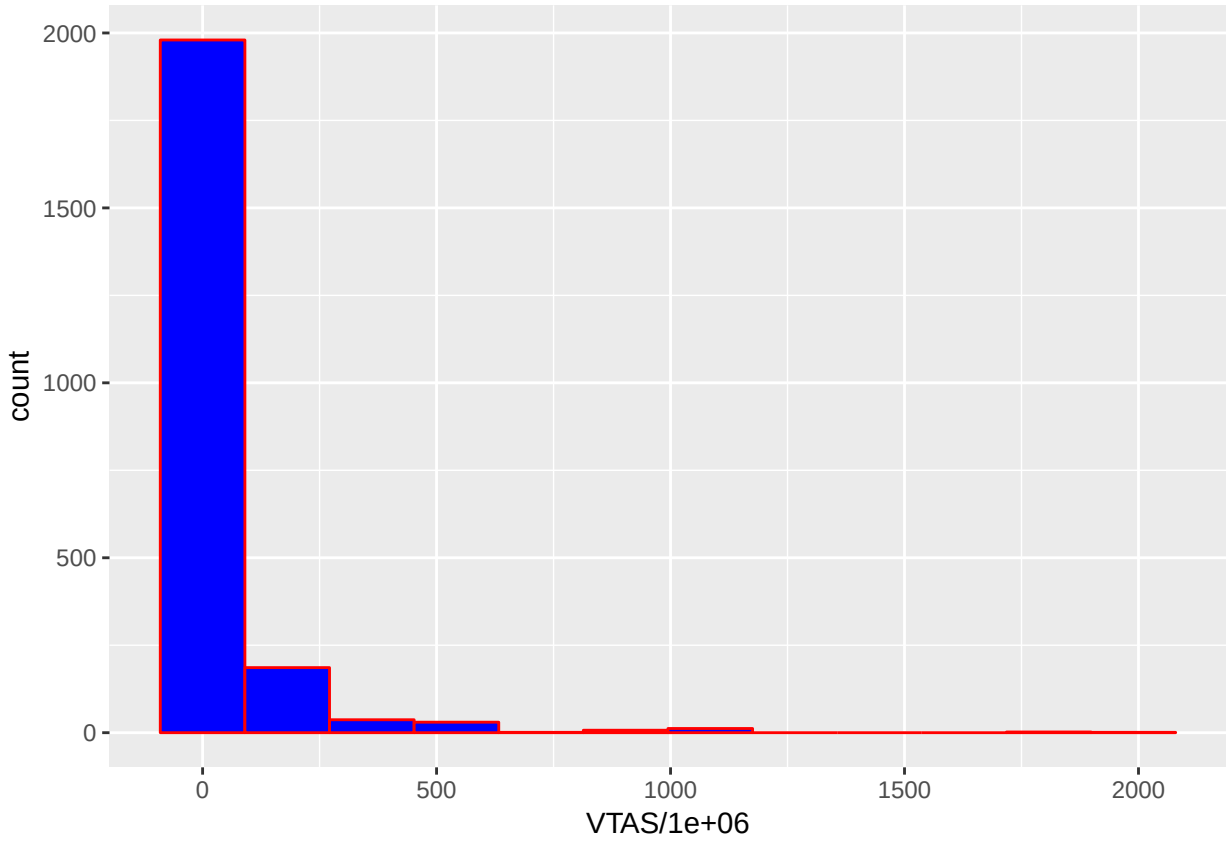

Figura 2.4: Histograma de las Ventas

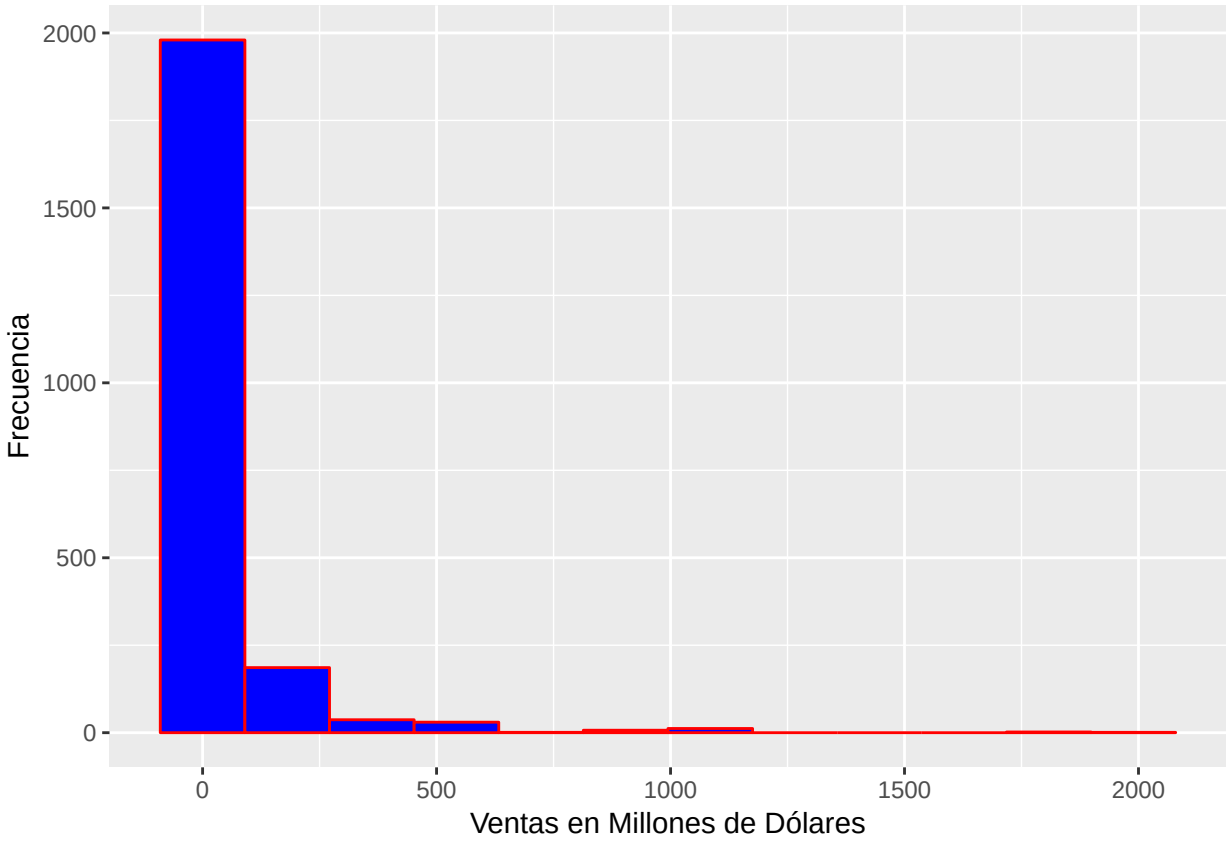

Figura 2.5: Histograma de las Ventas con Etiquetas en los Ejes

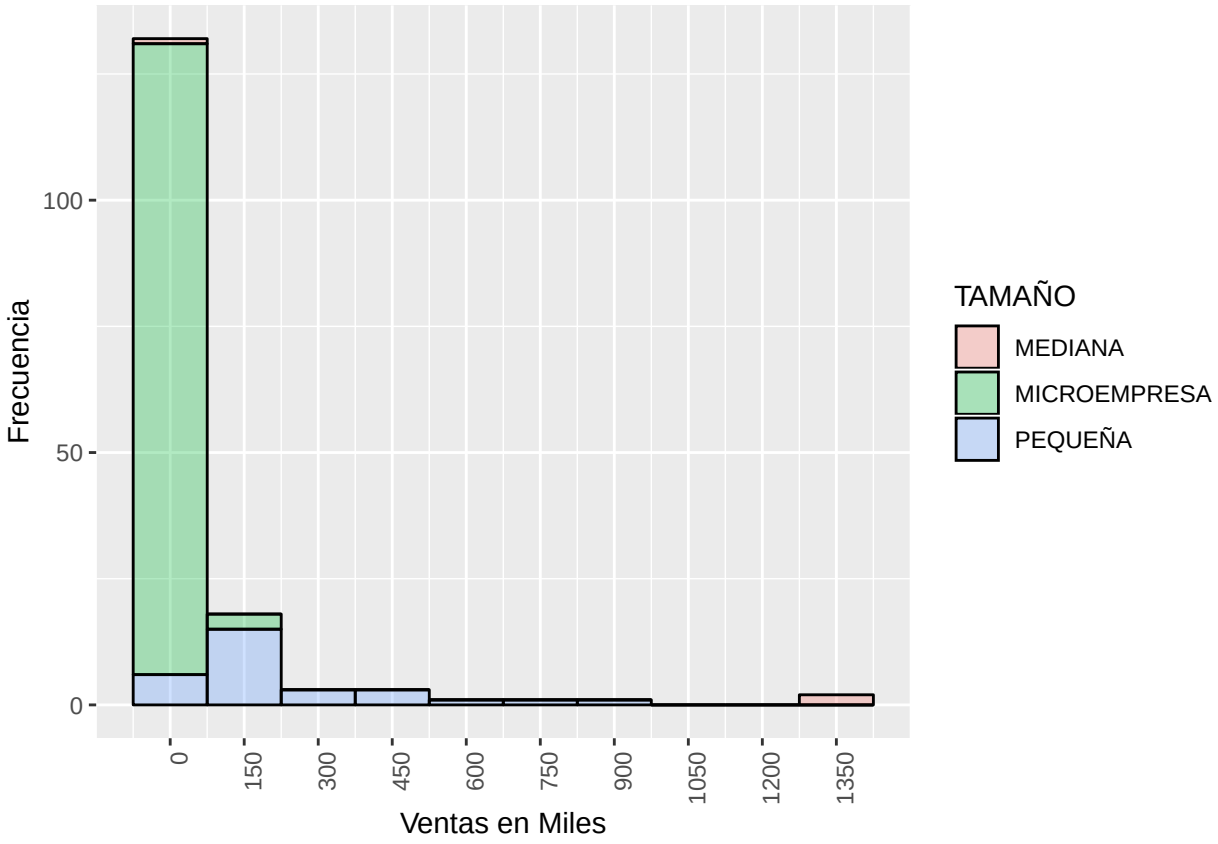

Figura 2.6: Histograma de las Ventas de Acuerdo al Tamaño de la empresa

Se puede observar en la 2.6 que algunas empresas medianas tienen mayores ventas que el resto de empresas. Una mejor forma de comparar la distribución de una variable de acuerdo a otra variable es usar los diagramas de caja que serán discutidos en profundidad en la sección 2.9.2.

2.9.2 Diagramas de Caja y valores atípicos

En la 2.6 se pretendía mostrar la distribución de las ventas de acuerdo al tamaño de la empresa. Sin embargo el histograma no mostraba claramente la distribución de acuerdo al tamaño de la empresa. una alternativa es usar un diagrama de caja.

Un diagrama de caja está formado por 5 valores que lo resumen, estos 5 valores se muestran en la figura 2.7. La distancia entre el primer y el tercer cuartil se la conoce como rango intercuartílico (IQR, por sus siglas en inglés). El límite superior es igual al tercer cuartil más 1.5 veces el rango intercuartílico, valores mayores a esta cantidad se consideran valores atípicos. Mientras que el límite inferior es igual al primer cuartil menos 1.5 veces el rango intercuartílico y valores menores a esta cantidad se consideran valores atípicos.

$$IQR = Q_3 - Q_1 \tag{2.13}$$

$$LS = Q_3 + 1.5IQR \tag{2.14}$$

$$LI = Q_1 - 1.5IQR \tag{2.15}$$

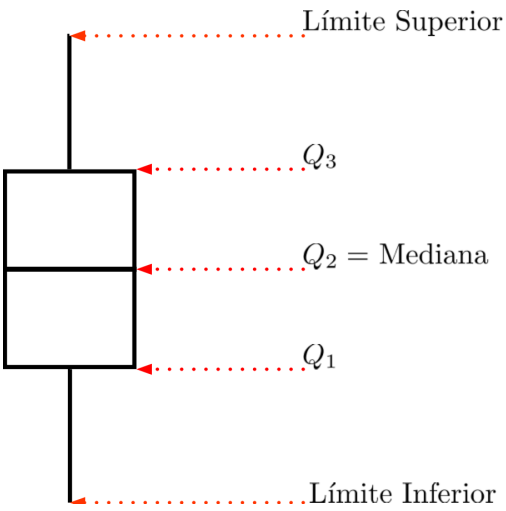

Figura 2.7: Partes de un Diagrama de Caja

En la figura 2.8 se observan los diagramas de caja de las ventas según el tamaño de la empresa. Se puede notar que existen diferencias entre las ventas de las empresas medianas, las microempresas y las pequeñas. El 50% de las empresas medianas vende más de 1250000, mientras que todas las microempresas venden menos de 250000. Las empresas pequeñas que venden más de 500000 son atípicas, mientras que en las empresas medianas no se presentan valores atípicos.

```
ggplot(rank2018, aes(TAMAÑO, VENTAS/1000)) +
  geom_boxplot() + xlab("Tamaño de las empresas") +
  ylab("Ventas en Miles de Dólares")
```

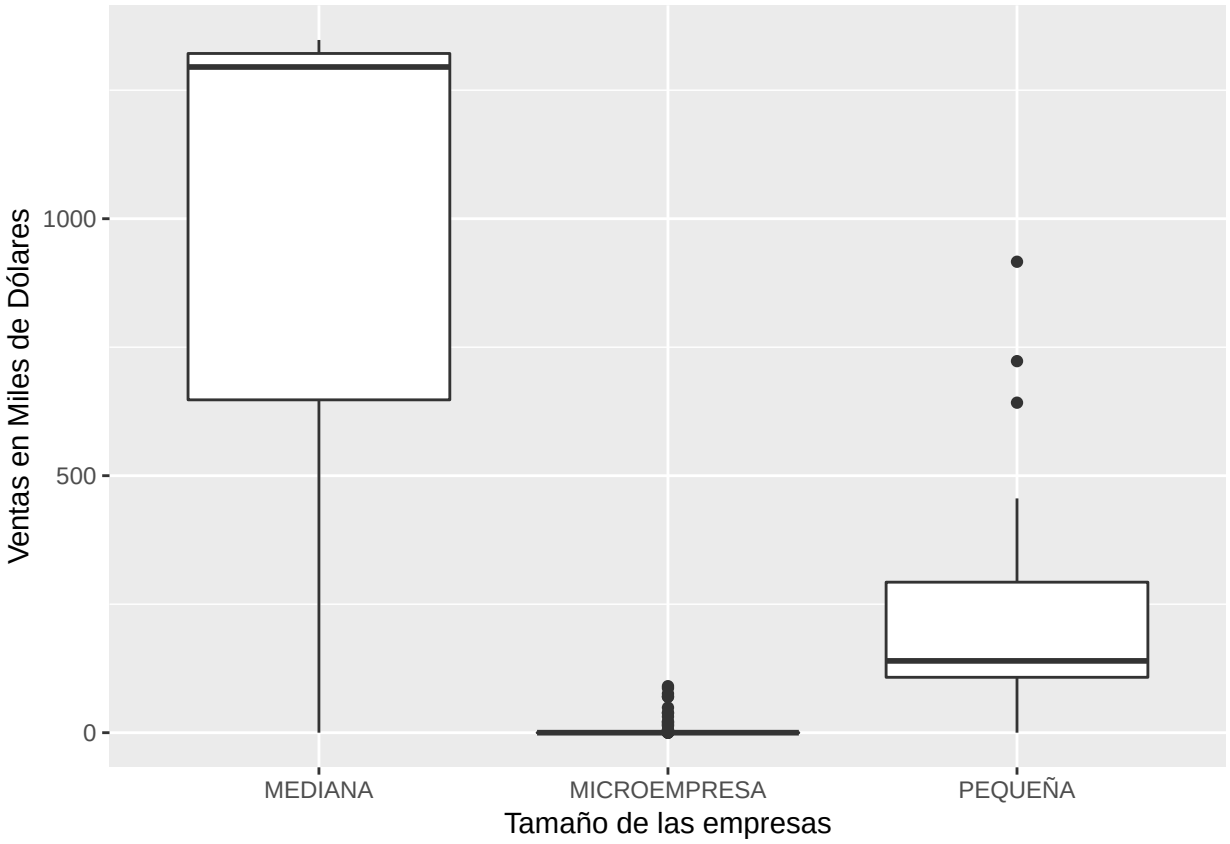

Figura 2.8: Diagrama de Caja de las Ventas según el Tamaño de la empresa

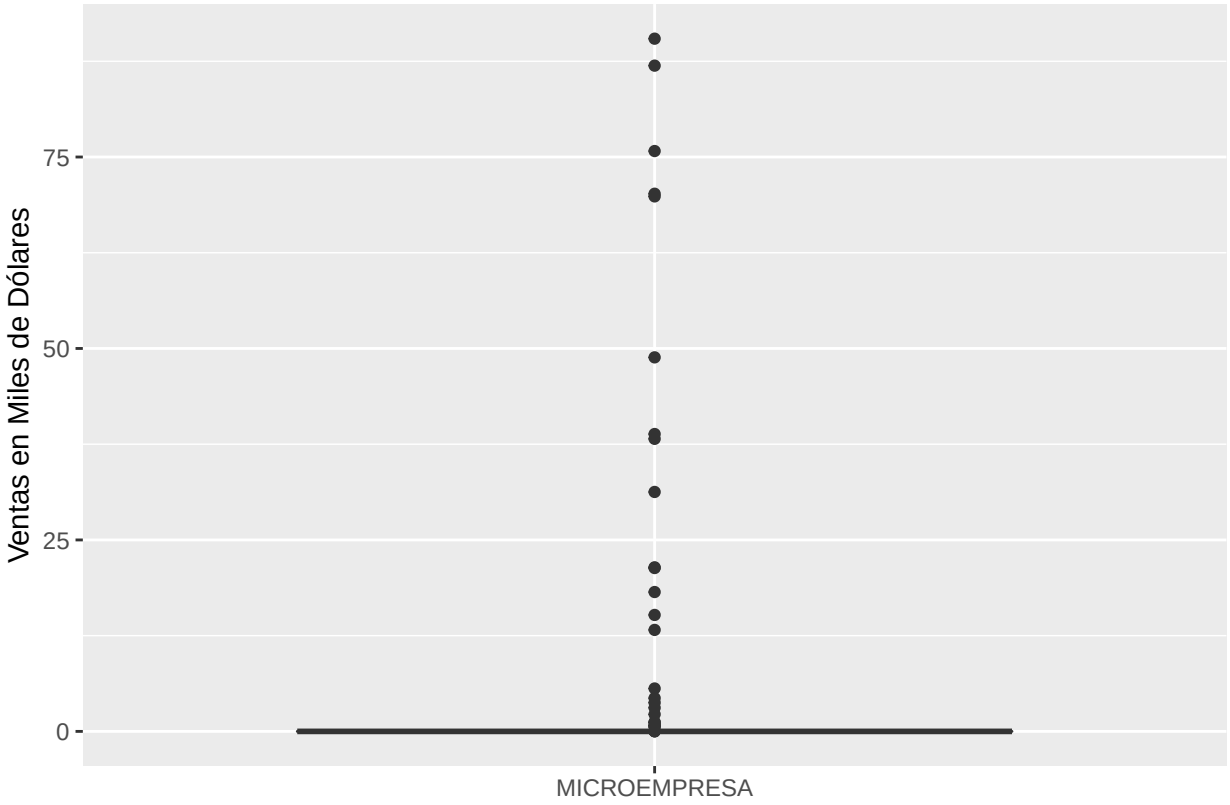

Figura 2.9: Diagrama de Caja de las Ventas de las Microempresas

Si se quisiera analizar con mayor detalle las microempresas se podría seleccionar solo las empresas con este tamaño y elaborar el diagrama de caja correspondiente, para lograr esto se utiliza la función `subset(df, cond)`. donde `df` corresponde al *data frame* usado y *cond* a la regla que deben cumplir los datos a ser analizados.

```
ggplot(subset(rank2018, TAMAÑO == "MICROEMPRESA"), aes(TAMAÑO, VENTAS/1000)) +
  geom_boxplot() + xlab("") +
  ylab("Ventas en Miles de Dólares")
```

Capítulo 3

Pruebas de Hipótesis

3.1 Intervalos de Confianza

3.2 Pruebas de hipótesis para la media

3.3 Pruebas de hipótesis para la proporción

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

Wilkinson, L. (2005). *The Grammar of Graphics*. Springer-Verlag, New York, 2nd edition. ISBN 978-0-387-28695-2.