Análisis Estadístico de Datos Financieros con R

Oswaldo Navarrete Carreño

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 datos 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 **razón**. 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 Otros conceptos importantes

Existen algunos conceptos que son importantes y que se deben tener presentes en el análisis estadístico de datos.

- **Población**: una población es el conjunto de todos los sujetos u objetos de interés en una investigación o análisis. Por ejemplo si se desea analizar la intención de voto en una ciudad para las próximas elecciones seccionales, la población serían todas las personas en edad de votar empadronadas en la ciudad.
- **Muestra**: es la parte de la población que es analizada. Sigamos con el ejemplo de la intención de voto, aunque el investigador quisiera no puede acceder a toda la población ya sea por cuestiones de tiempo o dinero y por esta razón debe tomar una parte de la población. La muestra debe representar lo mejor posible a la población. La parte de la estadística que comprende los métodos estadísticos para obtener muestras representativas de una población se llama *muestreo*
- **Parámetro**: un parámetro es una cantidad numérica que caracteriza a una población.
- **Estadístico**: un estadístico es una cantidad numérica que caracteriza a una muestra.

2.4 Primeros pasos en R

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

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 `=`.

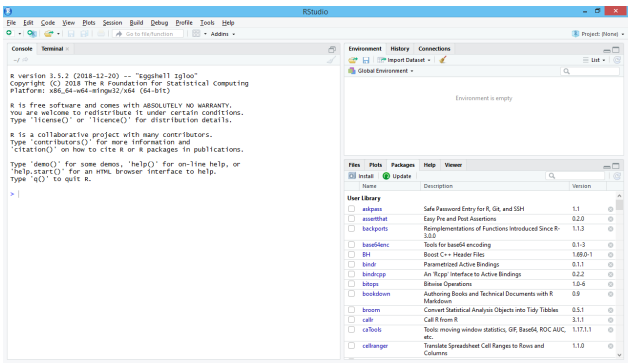

Figura 2.1: Ventana de RStudio

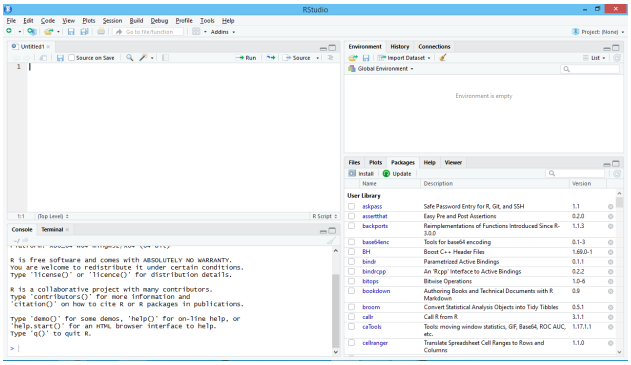

Figura 2.2: Ventana de RStudio con Script

En el código que se muestra a continuación en la variable `x` almacenaremos un conjunto de 8 observaciones, es importante observar que la lista de datos es ingresada con el comando `c(elem1,elem2,elem3,...)` si los elementos de la lista fueran cadenas de texto o caracteres cada elemento se encierra entre comillas `" "`:

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

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 existen dos formas la primera es fijar un directorio de trabajo y la segunda es crear un proyecto. Un directorio de trabajo es el espacio donde deben estar guardados nuestros archivos en el formato que sea de nuestra preferencia. Una forma de fijar el directorio de trabajo es desde la barra de menú escoger las opciones **Session, Set Working Directory, Choose Directory** o desde el teclado con la combinación **Ctrl+Shift+H**, o con la función `setwd("rutadelarchivo")`.

Sin embargo la mejor forma de trabajar, por la experiencia de los autores, es crear un proyecto. Para crear un proyecto primero debemos presionar **File, New Project** como se ve en la figura 2.3

Si no hemos creado la carpeta donde reposarán nuestros archivos, escogemos la primera opción **New Directory**. Pero si los archivos a ser analizados ya están en una carpeta escogemos la segunda opción **Existing Directory**, en este caso ya tenemos los archivos en una carpeta llamada *AEDFR* por lo que escogemos la segunda opción. Las opciones se aprecian en la figura 2.4

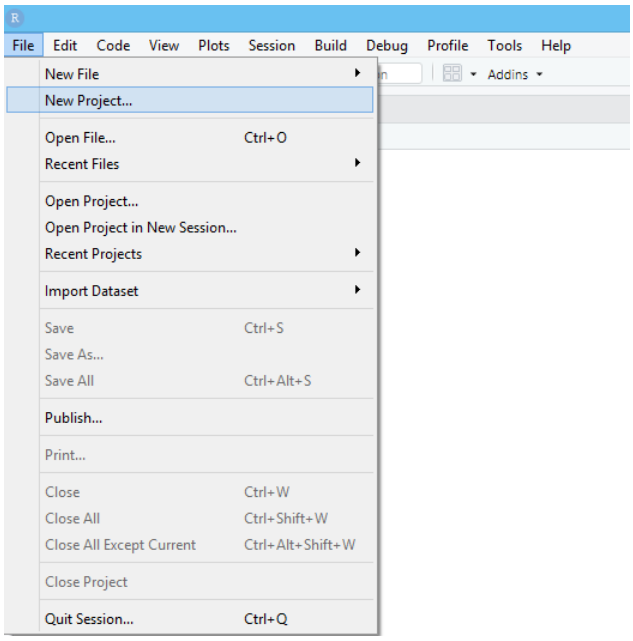

Figura 2.3: Nuevo Proyecto. Paso 1

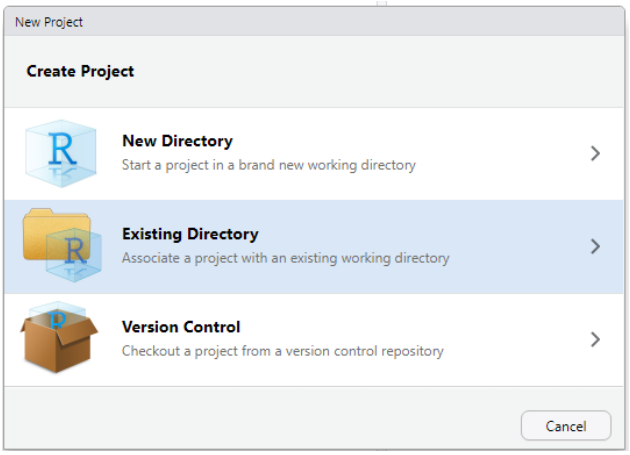

Figura 2.4: Nuevo Proyecto. Paso 2

Luego escogemos la ubicación de la carpeta donde reposan nuestros archivos dando clic en el botón **Browse**. Finalmente escogemos **Create Project** como se aprecia en la figura 2.5.

Finalmente se crea nuestro proyecto, la ventana de Rstudio ahora en la parte superior derecha ahora muestra el nombre de nuestro proyecto como se puede ver en la figura 2.6

En este primer ejercicio trabajaremos con el archivo *cap2_big4_size.csv*. Los datos serán guardados en una variable llamada `big4size`, 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 o no 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 operativo 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.

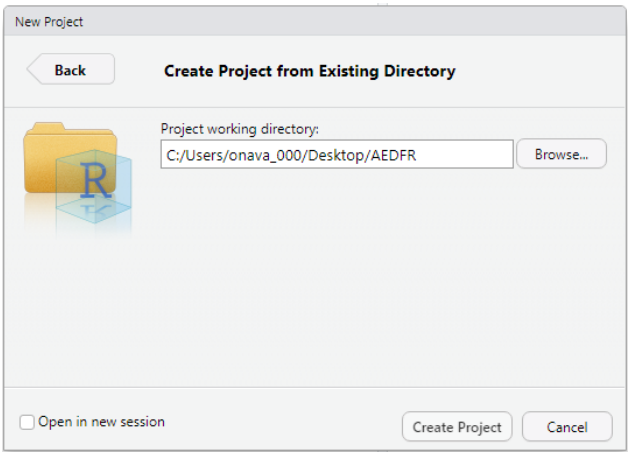

Figura 2.5: Nuevo Proyecto. Paso 3

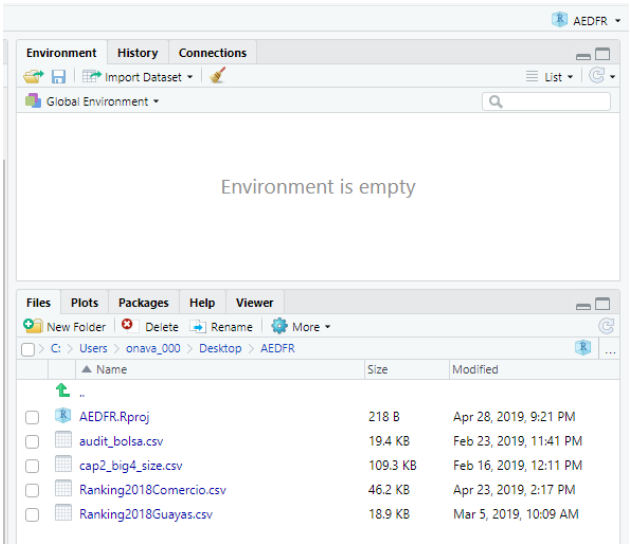

Figura 2.6: Nuevo Proyecto. Paso Final

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.

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

Sin más preámbulos, empecemos a trabajar. Recapitulando nuestro flujo de trabajo es:

1. Configurar el directorio de trabajo o crear un proyecto
2. Cargar el archivo indicado.
3. Finalmente usamos la función `str()`, que nos permite obtener la descripción de la estructura de los datos.

```
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)` (por el inglés *integer*) 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 aprenderemos 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` es decir una variable cuantitativa continua porque permite el uso de decimales.
- `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 como '`ggplot2`, `dplyr` y otros una descripción detallada de estos paquetes puede ser encontrada en https://www.rstudio.com/products/rpackages/. Trabajaremos también con paquetes desarrollados por otros colaboradores de la comunidad de usuarios de R.

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 `big4size` 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 tres nuevas variables. En la próxima sección seguiremos trabajando con el mismo conjunto de datos.

2.5 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.

2.5.1 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 expresión $\sum_{i=1}^{n}$ se interpreta como la suma desde el primer hasta el último elemento del conjunto de datos.

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.

2.5.2 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.

2.5.3 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.

2.5.4 ¿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.

2.5.5 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
```

2.5.5.1 ¿Cuándo usar `mutate()` y cuándo usar `summarise()`?

Note que cuando usamos la función `summarise()` creamos nuevas variables en el conjunto de datos, al igual que cuando se usa la función `mutate()` la principal diferencia entre `summarise()`y `mutate()` es que la primera resume los datos es decir devuelve un nuevo conjunto de datos con menos filas, mientras que `mutate()` devuelve el conjunto de datos con el mismo número de observaciones es decir con el mismo número de filas.

2.6 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.

2.6.1 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.

2.6.2 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 `big4size` y se llama `ACTIVOS` por lo que para llamar la variable desde la función escribimos `big4size$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, ... )` como ya lo habíamos indicado antes.

```
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 3 184 669. 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
```

2.7 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 iguales $\bar{X} = \bar{Y} = 5$. En la figura 2.7 se han graficado con color rojo los puntos del conjunto X y de color celeste

los puntos del conjunto Y. Se observa que los valores del conjunto Y están más dispersos que los valores del conjunto X, es fácil observar que los valores del conjunto X están más cercanos a la media. En esta sección se discute las formas existentes para cuantificar la dispersión.

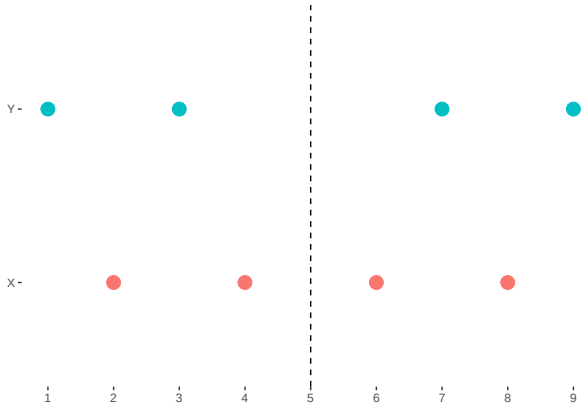

Figura 2.7: Conjuntos graficados

2.7.1 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:

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

2.7.2 Varianza

La dispersión en un conjunto x se puede entender como una medida de la distancia que tiene cada dato x_i a la media de los datos. En el caso del conjunto X descrito al inicio de esta sección se puede verificar que para cada dato la distancia del dato a la media $(x_i - \bar{x})$ es:

- $2 - 5 = -3$
- $4 - 5 = -1$
- $6 - 5 = 1$
- $8 - 5 = 3$

Sin embargo si sumamos estos valores el resultado es 0. Para cualquier conjunto de datos X se verifica que $\sum_{i=1}^{n} x_i - \bar{x} = 0$ por esta razón para calcular la dispersión se trabaja con la distancia cuadrática $(x - \bar{x})^2$.

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 muestra se puede calcular la varianza muestral:

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

2.7.3 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.

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

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

2.7.4 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
```

2.8 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.

2.8.1 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`. Recordemos que la función `mutate( )` sirve para crear nuevas columnas, en este caso se crea la columna porcentaje.

```
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 2.2: 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

FIRMA	Frecuencia	Porcentaje
HERRERA CHANG 6 ASOCIADOS	1	0.47
NGV	1	0.47

En la tabla 2.2 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 2.4.

2.8.2 Variables Cuantitativas

Una tabla de frecuencias para variables cuantitativas 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 2.4.

```
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 2.3: 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 2.3 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)
```

2.9 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.

```
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 2.4: 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 2.4 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 2.5 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 2.5: 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

2.10 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).

2.10.1 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 2.8 se muestra el histograma de la variable ventas en millones de dólares del archivo *cap2_big4_size.csv* ya descrito en la sección 2.4, 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.

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

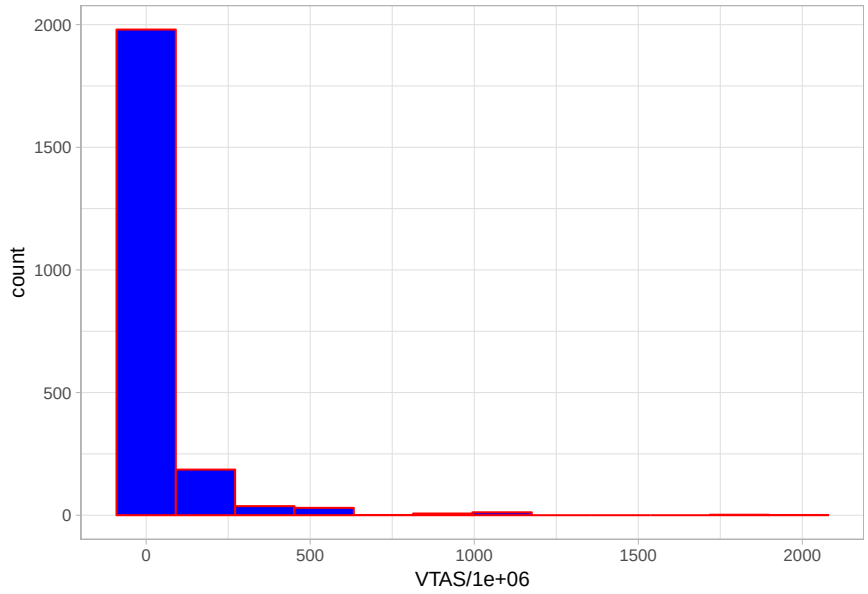

Figura 2.8: Histograma de las Ventas

Para configurar las etiquetas de los ejes podemos añadir las opciones `xlab( )` y `ylab( )`. En la figura 2.9 se aprecia el histograma con las etiquetas de los ejes añadidos.

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

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 2.10 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`

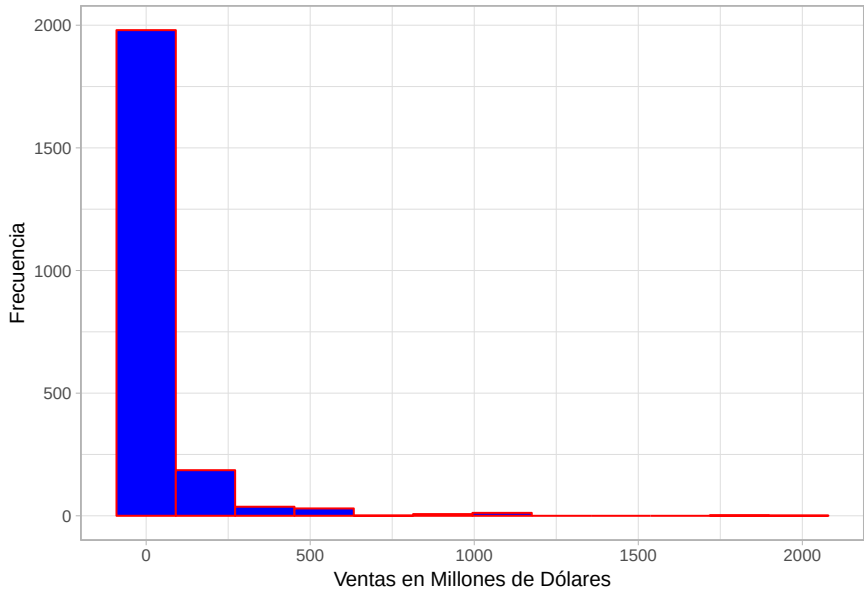

Figura 2.9: Histograma de las Ventas con Etiquetas en los Ejes

```
= 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),
        panel.background = element_rect(fill="white")) +
  xlab("Ventas en Miles") + ylab("Frecuencia")
```

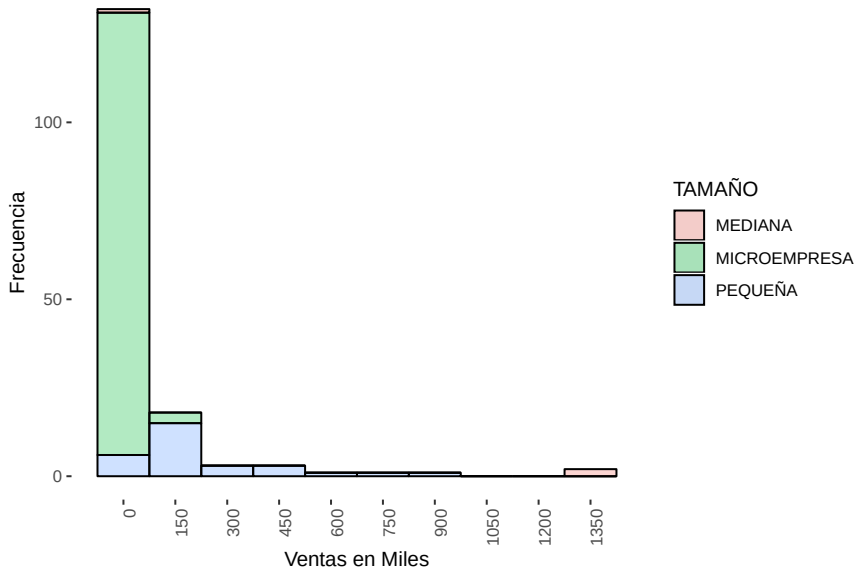

Figura 2.10: Histograma de las Ventas de Acuerdo al Tamaño de la empresa

Se puede observar en la 2.10 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 2.10.3.

2.10.2 Diagrama de barras

Los histogramas se usan para variables cuantitativas, mientras que los gráficos o diagramas de barras se utilizan para variables cualitativas. Al igual que los histogramas los diagramas de barras se elaboran para las frecuencias absolutas o las relativas.

Para ejemplificar la elaboración de diagrama de barras, vamos a trabajar con el conjunto de datos`Ranking2018Comercio.csv` que contiene una muestra de 507 empresas de Ecuador dedicadas al comercio.

```
rank2018com = read.csv("Ranking2018Comercio.csv",header = T,sep=";",dec=",")
str(rank2018com)
```

```
## 'data.frame':    507 obs. of  12 variables:
##  $ EXPEDIENTE: int  11069 103998 23140 165341 49377 93937 64400 134199 13963 95267 ...
##  $ TIPO      : Factor w/ 3 levels "ANÓNIMA","RESPONSABILIDAD LIMITADA",..: 1 1 2 1 1 2 1 1 2 2 ...
##  $ ACTIVIDAD : Factor w/ 112 levels "G4510.01","G4520.01",..: 69 69 71 69 42 1 4 26 60 76 ...
##  $ REGIÓN    : Factor w/ 4 levels "COSTA","GALAPAGOS",..: 4 1 1 1 4 4 4 1 4 4 ...
##  $ PROVINCIA : Factor w/ 16 levels "AZUAY","BOLIVAR",..: 12 7 5 7 12 14 12 7 12 1 ...
##  $ TAMAÑO    : Factor w/ 4 levels "GRANDE","MEDIANA",..: 1 1 1 1 1 1 1 1 2 1 ...
##  $ SECTOR    : Factor w/ 2 levels "MERCADO DE VALORES",..: 2 2 2 2 2 1 2 2 2 2 ...
##  $ EMPLEADOS : int  560 145 65 108 65 84 127 45 72 20 ...
##  $ ACTIVO    : num  19013933 16498975 13759911 13015775 12664280 ...
##  $ PATRIMONIO: num  5441046 11084842 1854108 4353938 4547432 ...
##  $ VENTAS    : num  27523262 54939461 16703973 20409277 19232071 ...
##  $ UTILIDAD  : num  2935434 1180875 229199 81757 -900404 ...
```

El conjunto de datos tiene 507 observaciones con 12 variables. Las variables presentes en este conjunto de datos son:

1. `EXPEDIENTE` variable cualitativa que almacena el número de expediente asignado por la Superintendencia de Compañías.
2. `TIPO` es una variable cualitativa con un nivel de medición nominal que almacena el tipo de compañía. En esta variable existen 3 niveles.
3. `ACTIVIDAD` variable cualitativa con nivel de medición nominal que almacena la actividad económica de la empresa, estas actividades se rigen por un catalógo dado por la superintendencia.
4. `REGIÓN` variable cualittativa con nivel de medición nominal que almacena la región del país a la que pertenece la empresa.
5. `PROVINCIA` variable cualitativa con nivel de medición nominal que almacena la provincia a la que pertenece la empresa.
6. `TAMAÑO` variable cualitativa con nivel de medición ordinal que almacena el tamaño de la empresa.
7. `SECTOR` variable cualitativa con nivel de medición ordinal que indica el sector al que pertenece la empresa.
8. `EMPLEADOS`variable cuantitativa discreta que almacena el número de empleados de la empresa.
9. `ACTIVO` variable cuantitativa continua que almacena el valor de los activos en libros de la empresa.
10. `PATRIMONIO` varaible cuantitativa continua que almacena el valor del patrimonio en libros de la empresa.
11. `VENTAS` varaible cuantitativa continua que almacena el valor de las ventas en libros de la empresa.
12. `UTILIDAD` varaible cuantitativa continua que almacena el valor de la utilidad en libros de la empresa.

Como hemos dicho anteriormente los diagramas de barra se utilizan para variables cualitativas. Vamos a comenzar elaborando una tabla de frecuencias de las empresas por región.

```
tabla_reg <- rank2018com %>%
  group_by(REGIÓN) %>%
  summarise(Frecuencia=n()) %>%
  mutate(Porcentaje = round(100*Frecuencia/sum(Frecuencia),2)
  ) %>%
  arrange(desc(Porcentaje))
print(tabla_reg)
```

```
## # A tibble: 4 x 3
##   REGIÓN    Frecuencia Porcentaje
##   <fct>          <int>      <dbl>
## 1 COSTA            368      72.6
## 2 SIERRA           133      26.2
## 3 ORIENTE            4       0.79
## 4 GALAPAGOS          2       0.39
```

Para hacer diagramas de barras se utiliza `geom_bar()` en este caso como deseamos hacer un diagrama de barras de las frecuencias dentro de `geom_bar()` esvribimos `stat="count"`. En la figura 2.11 se observa el gráfico de barras de las empresas por región.

```
ggplot(rank2018com, aes(x=REGIÓN)) +
  geom_bar(stat = "count",col="black",fill="white") +
  xlab("") + ylab("Frecuencia")
```

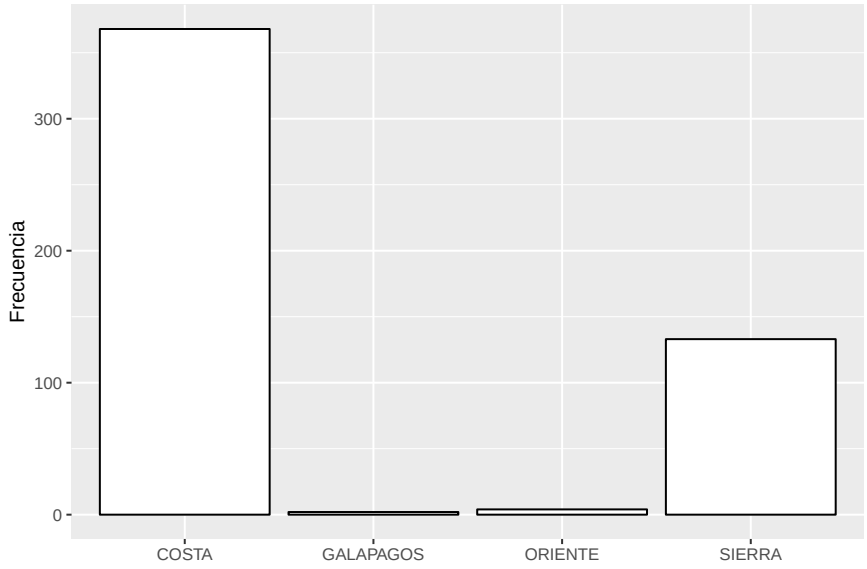

Figura 2.11: Gráfico de Barras de las empresas por Región

En la figura se observa que las regiones están en orden alfabético. Supongamos que se desea ordenar las regiones en el siguiente orden:

1. Costa
2. Sierra
3. Oriente
4. Galápagos

Para lograr este objetivo, antes de hacer el diagrama de barras debemos reordenar las regiones. Como se muestra a continuación:

```
rank2018com$REGIÓN <- factor(rank2018com$REGIÓN,
                             levels = c("COSTA", "SIERRA", "ORIENTE", "GALAPAGOS"))
```

Luego de reordenar las regiones podemos volver a realizar el gráfico de barras pero esta vez obtendremos las regiones en el orden deseado como se observa en la figura 2.12:

```
ggplot(rank2018com, aes(x=REGIÓN)) +
  geom_bar(stat = "count",col="black",fill="white") +
  xlab("") + ylab("Frecuencia")
```

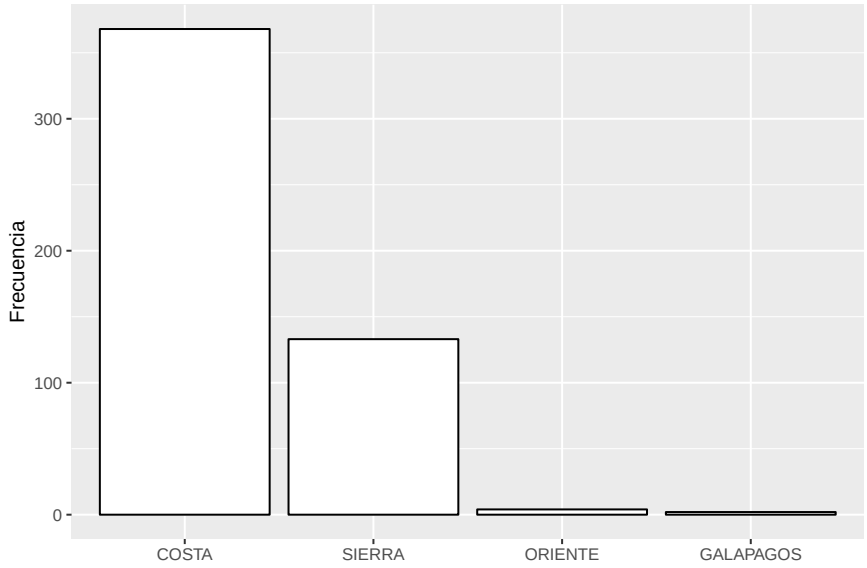

Figura 2.12: Gráfico de Barras de las empresas por Región (Ordenadas)

Supongamos ahora que queremos hacer un gráfico de barras de las empresas por región y por tamaño. Elaboremos primero una tabla de contingencia que contenga la información solicitada.

```
tama.reg = rank2018com %>%
  group_by(TAMAÑO, REGIÓN)%>%
  summarise(n=n())%>%
  spread(TAMAÑO, n) %>%
  replace(., is.na(.), 0)

print(tama.reg)
```

```
## # A tibble: 4 x 5
##   REGIÓN    GRANDE MEDIANA MICROEMPRESA PEQUEÑA
##   <fct>      <dbl>   <dbl>        <int>   <dbl>
## 1 COSTA          5      13          264      86
## 2 SIERRA         6      11           73      43
## 3 ORIENTE        0       0            3       1
## 4 GALAPAGOS      0       0            2       0
```

Para elaborar el gráfico agregamos la opción `fill = TAMAÑO` junto a `x=REGIÓN`

```
ggplot(rank2018com, aes(x=REGIÓN,fill=TAMAÑO)) +
  geom_bar(stat = "count",position = "dodge") +
  xlab("") + ylab("Frecuencia")
```

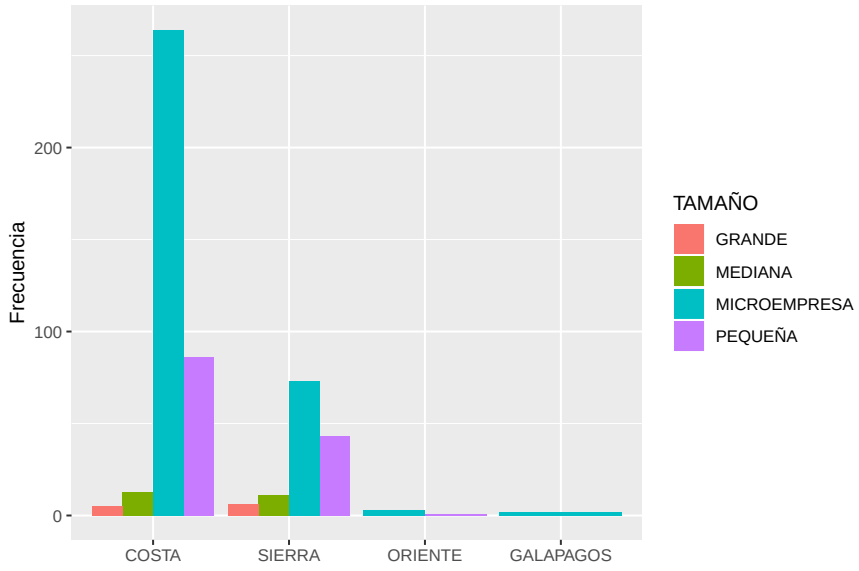

Figura 2.13: Gráfico de Barras de las empresas por Región y por Tamaño

2.10.3 Diagramas de Caja y valores atípicos

En la 2.10 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 2.14. 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.

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

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

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

En la figura 2.15 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") +theme_light()
```

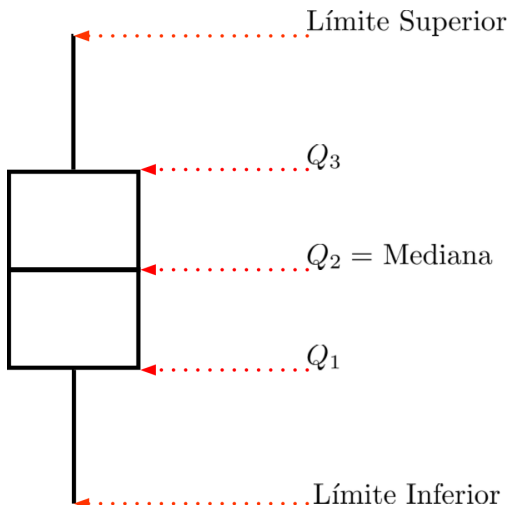

Figura 2.14: Partes de un Diagrama de Caja

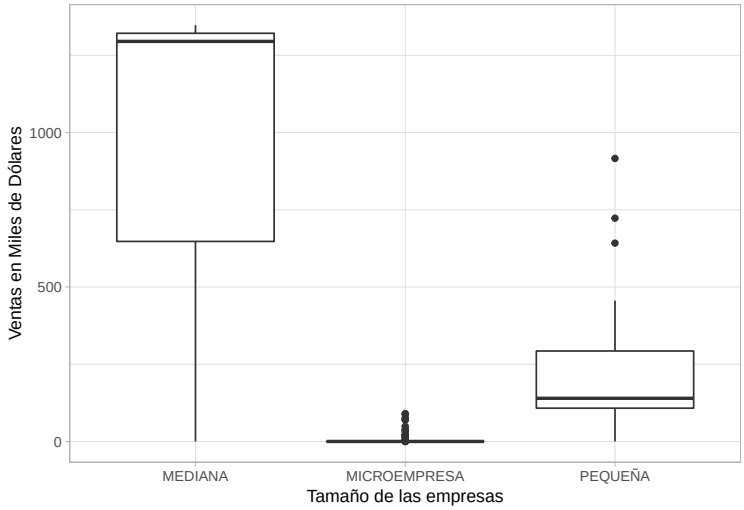

Figura 2.15: Diagrama de Caja de las Ventas según el Tamaño de la empresa

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") +theme_light()
```

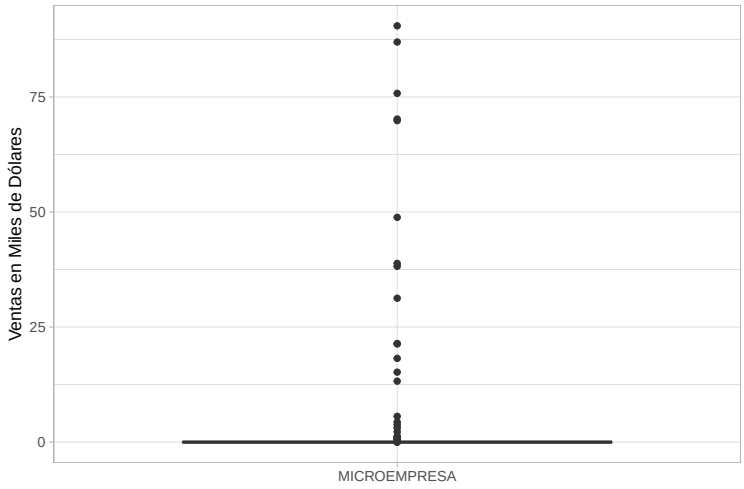

Figura 2.16: Diagrama de Caja de las Ventas de las Microempresas

Capítulo 3

Distribuciones de probabilidad

En este capítulo abordaremos las distribuciones de probabilidad que se necesitarán en las siguientes secciones. Es necesario para empezar, introducir el concepto de variable aleatoria.

3.1 Variable aleatoria y distribución de probabilidad

Consideremos un experimento donde lanzamos tres monedas y observamos los resultados. Podemos representar todos los eventos posibles:

- Cara en la primera moneda, cara en la segunda y cara en la tercera.
- Cara en la primera moneda, cara en la segunda y sello en la tercera.
- Cara en la primera moneda, sello en la segunda y cara en la tercera.
- Cara en la primera moneda, sello en la segunda y sello en la tercera.
- Sello en la primera moneda, cara en la segunda y cara en la tercera.
- Sello en la primera moneda, cara en la segunda y sello en la tercera.
- Sello en la primera moneda, sello en la segunda y cara en la tercera.
- Sello en la primera moneda, sello en la segunda y sello en la tercera

Pero podríamos también hacer una lista de estos eventos en una forma diferente, en lugar de definir cada evento indicando el resultado de cada moneda, podemos contar el número de caras o sellos. Por ejemplo si trabajamos con el número de sellos, los eventos ahora son:

- 0 sellos
- 1 sello
- 1 sello
- 2 sellos
- 1 sellos
- 2 sellos
- 2 sellos
- 3 sellos

El número de sellos recibe el nombre de **variable aleatoria**, las variables aleatorias pueden ser representadas con letras mayúsculas. Definamos a X como la variable aleatoria **número de sellos**. Además vamos a estar interesados en la probabilidad P de cada valor posible de X. En este caso los valores posibles de X son 0, 1, 2 y 3.

La probabilidad se la calcula de forma fácil como el número de resultados favorables sobre el número de resultados posibles. Por ejemplo si queremos obtener en este caso $P(X = 2)$, el número de resultados favorables es 3 porque en tres oportunidades se pueden obtener 2 sellos, mientras que el número de resultados posibles es 8. Es decir que $P(X = 2) = \frac{3}{8}$.

Formalmente podemos definir a una **variable aleatoria** como una regla o función que asigna un número a cada resultado de un experimento. Existen dos tipos de variables aleatorias las **discretas** y las **continuas**, las primeras se usan cuando los valores que puede tomar la variable son contables y las continuas se utilizan cuando los valores posibles de la variable son incontables.

Un ejemplo de variable aleatoria continua puede ser el tiempo que le puede tomar a una persona llenar su declaración de impuesto a la renta. Puede ocurrir que una persona tenga la información lista y le tome mínimo 30 minutos mientras que a otra persona sin la información lista le tome 5 horas máximo. Sea $X =$ tiempo en minutos para llenar una declaración con el valor mínimo 30 y el máximo de 300. Si intentamos contar el número posible de valores que puede tomar X, empezamos por el valor mínimo que es 30 y luego buscamos el valor que sigue de 30 ¿cuál es el valor que sigue de 30 en este caso? ¿31? ¿30.5? ¿30.05? ¿30.005? no podemos determinar el siguiente valor porque entre 30 y 30.005 por ejemplo existe una cantidad infinita de valores. Por lo tanto no podemos contar el número de valores posibles de X, y X es continua.

Una *distribución de probabilidad* es una tabla, fórmula o gráfico que describe los valores de una variable aleatoria y la probabilidad asociada a estos valores.

3.2 Funciones de Densidad de Probabilidad

Retomemos la variable aleatoria $X =$ tiempo en minutos para llenar una declaración. Si graficamos un histograma de frecuencias relativas, la altura de cada barra representa la proporción o porcentaje de valores en cada clase y la suma de todas las áreas es 1. Si el tamaño de la muestra aumenta, se puede reducir la longitud de cada clase y la altura va formando una curva más suave. La curva a la que se aproxima recibe el nombre de **curva de densidad** como se ve en la 3.1.

La densidad de la probabilidad puede ser descrita por una expresión matemática $f(x)$, que recibe el nombre de **distribución de probabilidad** o **función de densidad de probabilidad**. Las funciones de densidad de probabilidad de variables continuas cumplen con algunas propiedades:

- El área bajo la distribución de probabilidad es igual a 1
- La probabilidad de que X se encuentre en determinado intervalo (a, b) es igual al área bajo la curva entre los dos puntos a y b.
- $P(X = c) = 0$ para cualquier valor c para el que se encuentre definida la función de probabilidad.

3.3 Distribución de Probabilidad Normal

La distribución de probabilidad más usada para describir variables aleatorias continuas es la **distribución de probabilidad normal**. Esta distribución se la puede encontrar en variables como la altura de personas, los pesos, calificaciones, mediciones científicas, cantidad de lluvia, etc.

La distribución normal tiene la forma de una campana, como se observa en la figura 3.2

La función de densidad de probabilidad de la distribución normal es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3.1}$$

como se puede observar en la ecuación (3.1), la distribución normal tiene dos parámetros la media μ y la desviación σ cuando una variable aleatoria X tiene una distribución normal se escribe $X \sim \mathcal{N}(\mu, \sigma)$, y se observan las siguientes propiedades de acuerdo a estos parámetros.

- La distribución normal alcanza su máximo en la media, que se ubica en el centro de la curva. Es decir que la distribución normal es simétrica respecto a la media.

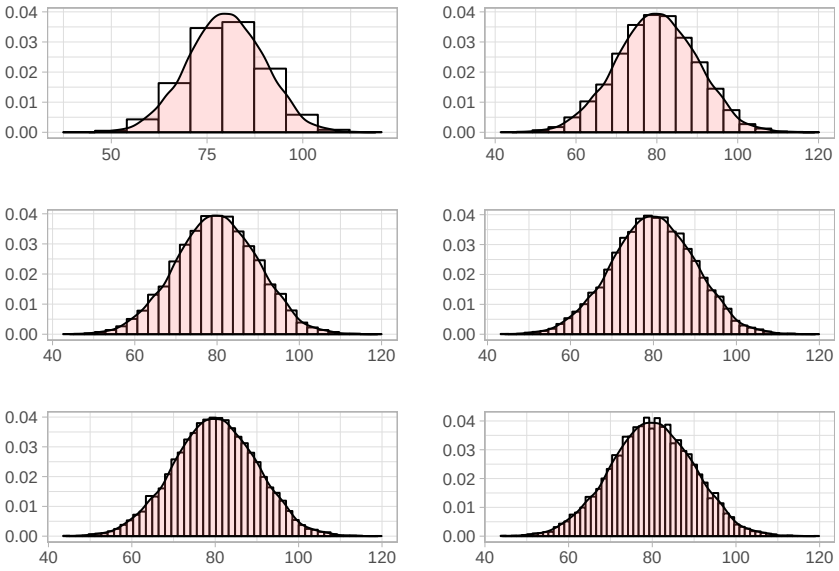

Figura 3.1: Curva de Densidad

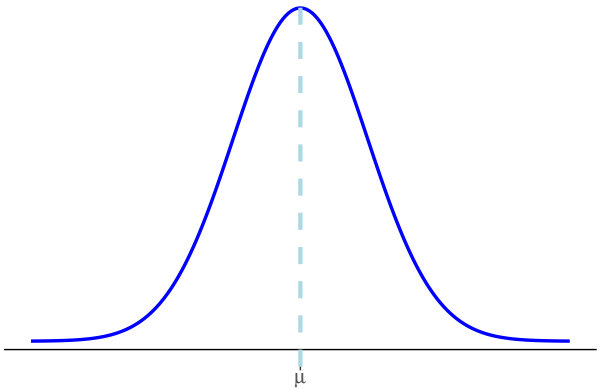

Figura 3.2: Distribución Normal

- La media es un parámetro de localización, en la figura 3.3 se observan tres distribuciones normales con la misma desviación pero diferentes medias.

- La desviación es un parámetro que afecta a la forma de la curva, a mayor desviación la curva se acerca hacia el eje de las X es decir la curva se "aplana". Y a menor desviación la curva se "estrecha". Este comportamiento se aprecia en la fiugra 3.4

- El porcentaje de valores en algunos intervalos que se usan comúnmente son:
  - 68.3% de los valores de una variable normal aleatoria se encuentran a más o menos una desviación de la media, como se muestra en la figura 3.5
  - 95.4% de los valores de una variable normal aleatoria se encuentran a más o menos dos desviaciones de la media, como se muestra en la figura 3.6
  - 99.7% de los valores de una variable normal aleatoria se encuentran a más o menos tres desviaciones de la media, como se muestra en la figura 3.7

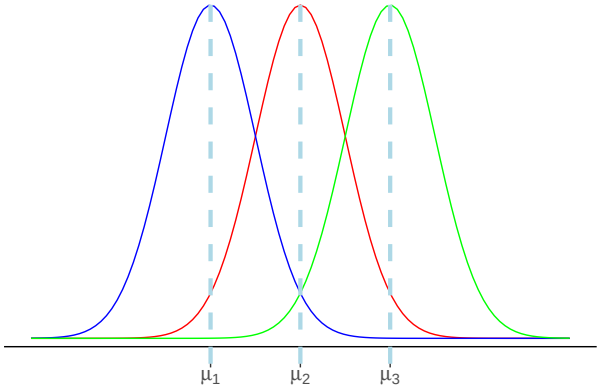

Figura 3.3: Diferentes distribuciones normales con diferentes medias

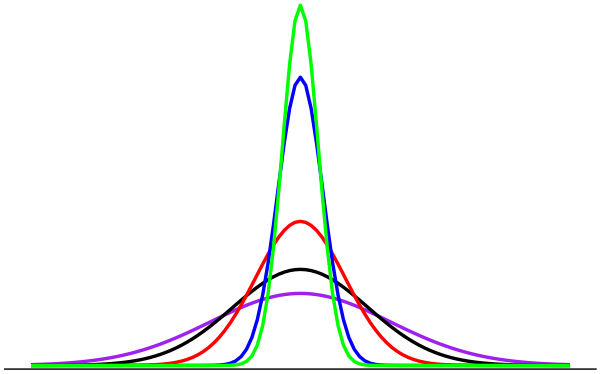

Figura 3.4: Diferentes distribuciones normales con diferentes desviaciones

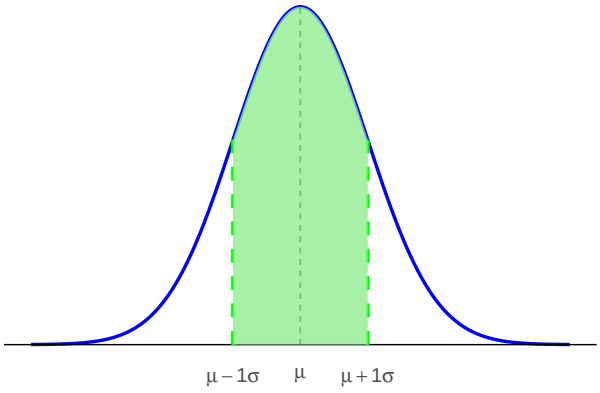

Figura 3.5: Porcentaje a 1 desviación de la media

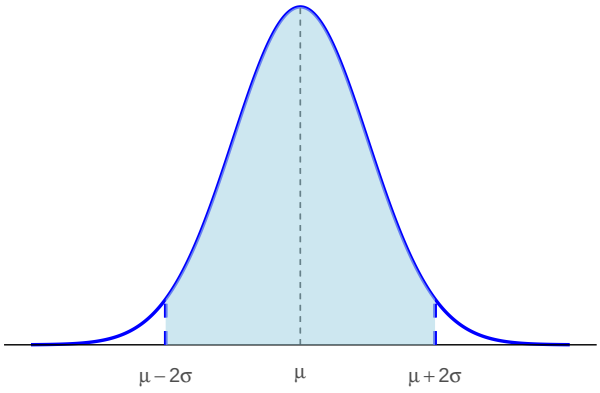

Figura 3.6: Porcentaje a 2 desviaciones de la media

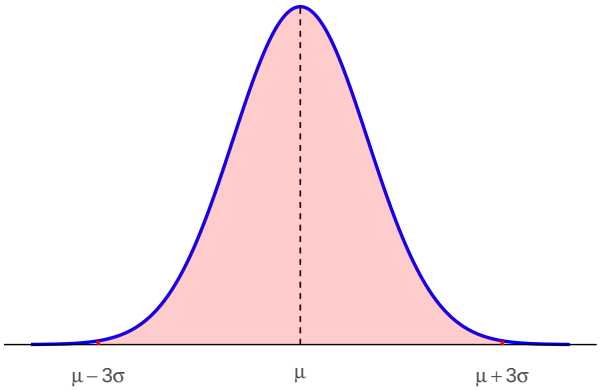

Figura 3.7: Porcentaje a 3 desviaciones de la media

3.3.1 Distribución Normal Estándar

Supongamos que el tiempo para llenar la declaración del impuesto a la renta se distribuye normalmente con media 80 y desviación 20 y que nos interesa determinar la probabilidad de que una declaración sea correctamente llenada entre 50 y 90 minutos. Formalmente si definimos a $X =$ Tiempo para llenar una declaración podemos decir que $X \sim \mathcal{N}(80, 20)$, nuestro interés es calcular $P(50 < X < 90)$ si hacemos el cálculo a mano debemos determinar el área bajo la función definida en la ecuación (3.1) entre 50 y 90 para $\mu = 80$ y $\sigma = 20$ esto requeriría de un trabajo fuerte desde el punto de vista matemático.

Otra solución sería tener una tabla que contenga las probabilidades de la distribución normal, sin embargo existe una cantidad infinita de distribuciones normales tomando en cuenta las combinaciones posibles de μ y σ. La solución es tener una distribución normal que represente a todas las distribuciones normales, dicha distribución recibe el nombre de **distribución normal estándar**.

La distribución normal estándar se caracteriza por tener media 0 y desviacion 1, y se representa con la letra Z. Dicho de otra forma $Z \sim \mathcal{N}(0, 1)$. En la figura 3.8 se presenta el gráfico de la normal estándar. Cualquier variable normal X se puede transformar a una normal estándar (estandarizar) realizando la operación $z = \frac{x - \mu}{\sigma}$. Una vez estandarizado se puede determinar la probabilidad deseada usando cualquier tabla de la normal estándar. En este texto no hacemos uso de tablas puesto que los cálculos de probabilidades los haremos utilizando R, sin embargo es importante que el lector aprenda o refresque la noción de la distribución

normal estandar ya que aparecerá constantemente en los siguientes capítulos.

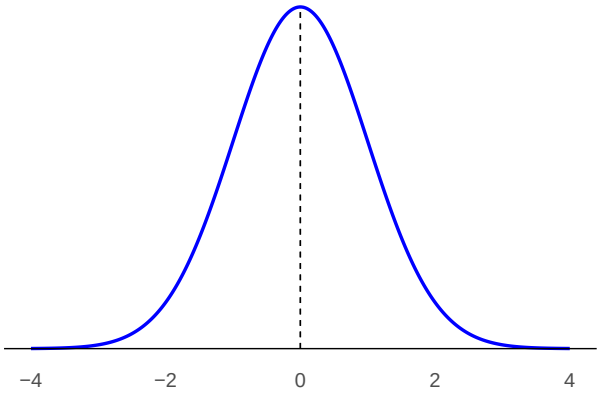

Figura 3.8: Normal Estándar

3.4 Distribución t de Student

La distribución t de Student fue obtenida por William S. Gosset en 1908, quien publicó sus hallazgos bajo el seudónimo "Student" y usó la letra t para representar la variable aleatoria, de aquí la **distribución t de Student** o t **de Student** simplemente. La función de densidad de la distribución t se define

$$f(t) = \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{t^2}{\nu}\right]^{\frac{\nu+1}{2}} \tag{3.2}$$

donde ν, (la letra griega *nu*) representa el parámetro de la distribución t llamado **grados de libertad** y Γ es la función Gamma cuya definición no será abordada en este texto. En la figura 3.9 se muestra el gráfico de dos distribuciones t de Student para dos valores diferentes de grados de libertad, la curva de color rojo representa una t con 3 grados de libertad y la de color verde con 9 grados de libertad. Se puede observar que la distribución t tiene una forma similar a la normal estándar porque es simétrica respecto a 0 (lo que implica que la media es 0) y tiene forma de campana.

En la figura 3.10 se muestra de color azul la curva de la normal estándar junto con las distribuciones t mostradas anteriormente se observa que las t tienen más probabilidad en las colas y menos en el centro. Además se puede observar que t se aproxima a z a medida que aumentan los grados de libertad, los grados de libertad están relacionados al tamaño muestral es decir que t converge a z para tamaños muestrales grandes. En el capítulo 4 se hace uso de esta distribución.

3.5 Cálculo de probabilidades y uso de la distribución normal y la t de Student en R

Todas las distribuciones que se manejan en R tienen cuatro funciones. Existe un nombre base, por ejemplo para la normal el nombre base es `norm`. La base es antecedida por cualquiera de las siguientes cuatro letras.

- `p` por *probabilidad*, es decir la función de distribución acumulada

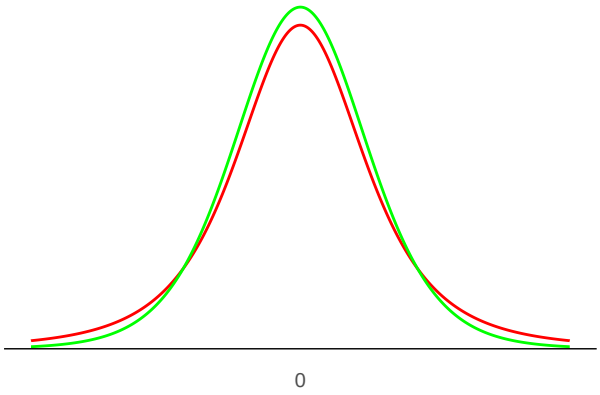

Figura 3.9: t de Student

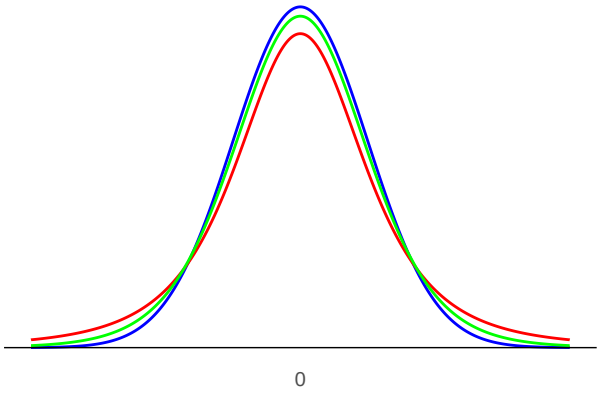

Figura 3.10: t de Student comparada con la Normal Estándar

- `q` por *cuantil*, en este caso se refiere a la función de distribución acumulada inversa es decir dado un valor de probabilidad nos devuelve el valor de la distribución que acumula esa probabilidad.
- `d` por *densidad*, es decir la función de densidad
- `r` por *aleatorio*, una variable aleatoria con la distribución especificada.

Para el caso de la distribución normal y la distribución t de Student entonces las funciones son:

- `pnorm`, `qnorm`, `dnorm` y `rnorm`
- `pt`, `qt`, `dt`, `rt`

3.5.1 Ejemplos

El cálculo de probabilidades con la distribución normal se hace con la función `pnorm`. La función `pnorm` se utiliza de la siguiente forma `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)` las opciones que van dentro de la función son:

- `q`: que es el valor para el cual se quiere determinar la probabilidad.
- `mean`: es el valor de la media de la distribución, el valor por defecto es 0 que corresponde a la media de la distribución normal estándar.
- `sd`: es el valor de la desviación de la distribución, el valor por defecto es 1 que corresponde a la desviación de la distribución normal estándar.

- `lower.tail`: se puede indicar "TRUE" o "FALSE", si es verdadero ("TRUE") se determina $P(X \leq q)$ y si es falso ("FALSE") se determina $P(X > q)$
- `log.p`: si es verdadero las probabilidades se dan en logaritmos.

En la sección 3.3.1 se puso un ejemplo sobre el tiempo de llenado de una declaración del impuesto a la renta. Recordemos que se partía de la suposición que el tiempo de llenado de la declaración del impuesto a la renta sigue una distribución normal con media 80 y desviación 20 minutos. y se quería calcular la probabilidad de que el llenado de una declaración se dé entre 50 y 90 minutos, es decir que deseamos calcular $P(50 \leq X \leq 90)$. El cálculo de la probabilidad se muestra a continuación.

```
pnorm(90,mean=80,sd=20)-pnorm(50,mean=80,sd=20)
```

```
## [1] 0.6246553
```

No hemos incluido las opciones `lower.tail` o `log.p` ya que son opciones en las que estamos usando los valores que por defecto tiene la función configurada. Si deseamos calcular la probabilidad que el tiempo de llenado de la declaración sea más de 100 minutos, para esto debemos indicar la opción `lower.tail = FALSE`.

```
pnorm(100,mean=80,sd=20, lower.tail = FALSE)
```

```
## [1] 0.1586553
```

Supongamos ahora que se quieren simular 100 valores que siguen la distribución normal con media 80 y desviación 20, utilizamos entonces la función `rnorm`, esta función se usa de la siguiente forma `rnorm(n, mean = 0, sd = 1)`

```
rnorm(100, mean = 80, sd=20)
```

```
##   [1]  47.60923  98.96942  67.24461  86.16940  67.04016  66.25843  61.04778
##   [8]  50.38047  98.94420  72.19053  69.68588  90.30176  62.54405  59.95936
##  [15]  82.32314  82.85486  59.47258  75.83318  77.84578  33.78631  94.24315
##  [22]  64.97883  81.35430  56.54594  99.09123  77.28837  87.03033  62.30110
##  [29]  47.07591  63.07407  86.09855  66.96359  63.28069  94.44924  71.94656
##  [36]  87.34025  57.46592  65.84012  87.93984  72.68598  53.39045  55.29303
##  [43]  67.72395  54.19637  80.45966  94.21219  85.25759  86.16999  95.79983
##  [50]  54.91902  73.21013  94.44568  89.48457  71.08272  75.01862  59.33797
##  [57]  66.44737  95.83850  90.92348  83.71054  78.17035 112.67491  95.60348
##  [64]  77.96828  77.34308  36.22008  53.64770  66.64198  72.19990  50.92120
##  [71]  55.96343  55.71079  52.48522  66.32572  98.44524  64.16898  81.52070
##  [78]  99.69316 101.73309  65.95563  73.38932  89.79295  53.62330  64.51062
##  [85]  56.90119  87.14268  80.07021  60.85577  77.94086  59.84116  77.76579
##  [92]  95.52738  82.81831 102.81758  70.26175  45.59517  70.96235  87.87413
##  [99]  82.72691  43.00896
```

Si queremos simular valores de una distribución t de Student utilizamos la función `rt`, que recibe los parámetros `rt(n, df, ncp)`. El parámetro `ncp` es la no centralidad. La distribución t que hemos revisado tiene su parámetro de no centralidad igual a 0.

```
rt(100, df = 19)
```

```
##   [1]  1.05793212 -1.91419617 -0.98003248  1.82127814 -0.49177192
##   [6] -1.60876208 -0.24666830 -0.61907163  0.89161411 -1.07121399
##  [11]  0.55777858 -0.34173324 -0.22608348  0.60352125 -0.43530532
##  [16] -2.70888929 -0.14397661  1.47236783  1.12144674  0.54959945
##  [21] -0.93454398  0.46496417  0.01606558  0.51325380 -0.50768504
##  [26]  1.23056132  1.17233938  0.45786037  0.71973992 -0.16469405
##  [31]  2.35267693  1.64599410  0.59061054 -0.32029886  1.61554201
##  [36]  0.23667190 -1.14439804 -0.34142114 -0.10384022 -1.17344169
##  [41]  0.16521510 -0.46672483  0.84081554 -0.13078948 -0.80756773
```

```
##  [46] -0.29228584  0.31121386 -1.11066753 -0.14827508  0.61646142
##  [51]  1.62607652  0.97964258  0.18385486 -1.40560155 -1.55718392
##  [56]  3.15454077  0.96139249  0.43710936  1.37441840 -0.25103471
##  [61] -0.74573823 -0.22579408 -0.28792816 -0.04133293 -1.75343816
##  [66] -1.11549807 -0.37535274  0.20485858 -0.87647931  0.49113814
##  [71]  2.23549101 -0.03244843  0.07789724 -0.07636843  0.33623236
##  [76] -0.54203596 -0.28451629  0.96882369 -0.57050798 -0.36881492
##  [81] -0.52990836  0.10764146 -0.35474714  0.22103691 -1.80767942
##  [86] -1.17944768  0.89669816 -0.71749452  0.32734466  0.09742948
##  [91]  0.12345667  0.64564900 -1.02498351 -1.16696937 -0.07550280
##  [96] -1.54658637 -0.20088762  2.39340654 -1.11400883 -1.37399043
```

Capítulo 4

Intervalos de Confianza y Pruebas de Hipótesis

En la sección 2.1 se habló de los típos de estadística, la estadística inferencial consiste de los métodos por medio de los cuales se puede hacer inferencias o generalizaciones sobre una población. La inferencia estadística se puede dívidir en dos grandes áreas: **estimación** y **pruebas de hipótesis**.

Imaginemos que se desea estimar el promedio de las ventas en miles de las empresas que son auditadas por firmas Big Four, sin embargo debemos recordar que en nuestro conjunto de datos no tenemos a todas las empresas sino a una muestra de las empresas, por lo que afirmar que el valor obtenido es el promedio de todas las empresas auditadas por Big Four es muy arriesgado. Sin embargo, podríamos dar un intervalo en el que posiblemente se encuentre el valor que deseamos estimar.

Ahora supongamos que usted como investigador quiere probar que el promedio de las ventas en miles de las empresas que son auditadas por una Big Four es mayor a las empresas que no son auditadas por una Big Four. Una primera aproximación para resolver este problema es realizar una gráfica que le muestre las ventas de acuerdo al tipo de empresa auditora.

Antes de elaborar el gráfico vamos a crear una nueva variable llamada `Big4`en la que si la variable ‘`BIG4` es igual a 1 la variable `Big4` tomará el valor de Sí, caso contrario tomará el valor de No.

```
big4size <- big4size %>%
  mutate(
    Big4 = ifelse(BIG4==1, "Sí","No")
  )
```

En las figuras 4.1 y 4.2 se observa el histograma y el diagrama de caja de las ventas de acuerdo a la firma auditora. Al observar las gráficas se puede afirmar que evidentemente el promedio de las ventas de las empresas auditadas por firmas Big Four es mayor, sin embargo en estadística no se puede confirmar o negar una afirmación con solo ver un gráfico.

```
ggplot(big4size, aes(x=VTAS/1000, fill=Big4)) +
  geom_histogram(alpha=0.3, color="black",bins=10, binwidth = 200000) +
  scale_x_continuous(breaks = seq(0,2000000,200000)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  xlab("Ventas en Miles") + ylab("Frecuencia")
```

```
ggplot(big4size, aes(Big4, VTAS/1000)) +
  geom_boxplot() + xlab("Tipo de Firma") +
  ylab("Ventas en Miles de Dólares")
```

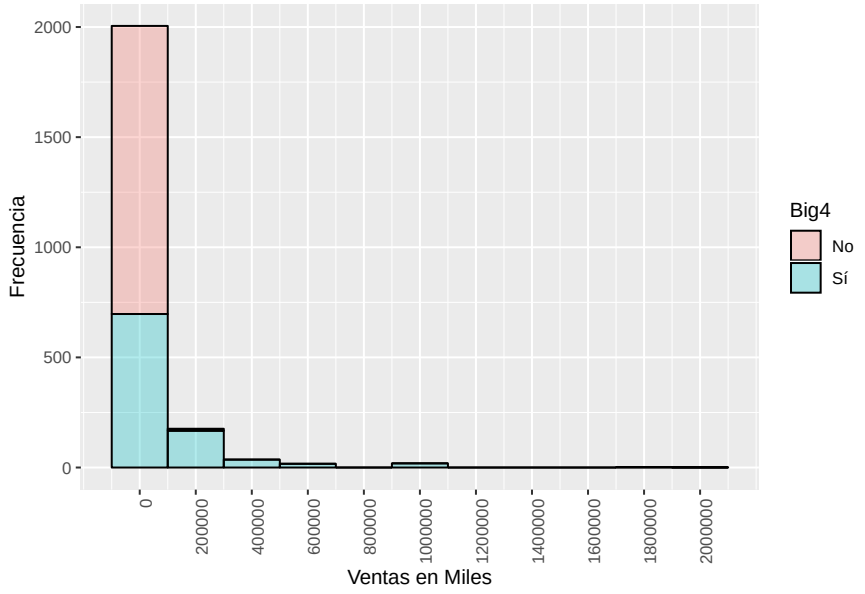

Figura 4.1: Histograma de las Ventas de Acuerdo al tipo de Firma Auditora

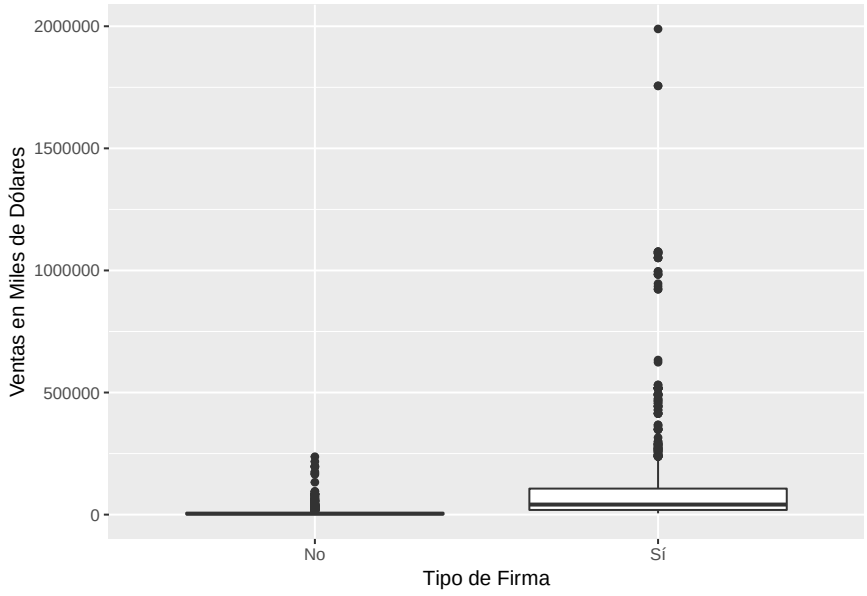

Figura 4.2: Diagrama de Caja de las Ventas de Acuerdo al tipo de Firma Auditora

Los dos problemas citados serán abordados y resueltos en este capítulo. El primero en la sección 4.1 y el segundo en la sección 4.2.

4.1 Intervalos de Confianza

Un *estimador puntual* de un parámetro θ es un número $\bar{\theta}$ de un estadístico Θ que puede ser considerado un valor que se aproxima a θ. Por ejemplo $\bar{x}$ del estadístico $\bar{X}$, calculado de una muestra de tamaño n es un estimador puntual del parámetro poblacional μ. Debido a que un estimador puntual es un simple número,

no da información por si solo sobre la precisión y la confiabilidad de la estimación.

En cualquier estimación de un parámetro habrá un error asociado, por ejemplo $\bar{X}$ no debe estimar μ con exactitud, pero se espera que no esté muy lejos del valor real. Lo que se espera de un estimador es que sea insesgado y eficiente.

La forma de un intervalo de confianza es:

$$\text{Estimador puntual} \pm \text{Margen de Error} \tag{4.1}$$

El estimador es el valor calculado a partir de la muestra para el parámetro desconocido. El *margen de error* es cuán preciso es nuestro cálculo, basados en la variabilidad del estimador, y de la confianza que tengamos en que el procedimiento detectará el valor real del parámetro de la población.

4.1.1 Interpretación de un intervalo de confianza

Un intervalo de confianza del $C\%$ indica que si construimos muchos de esos intervalos, entonces el $C\%$ de las veces el intervalo contendrá el valor real del parámetro.

Por ejemplo en la figura 4.3 se muestran 100 intervalos construidos con el 95% de confianza para el promedio poblacional, para construir cada intervalo se tomaron muestras de 100 elementos. Los intervalos de color celeste son los que contienen el valor real del parámetro, los intervalos de color rojo son los que no contienen el valor real del parámetro. El lector puede verificar que 5 intervalos es decir el 5% no contiene el valor real del parámetro lo que implica que el 95% de los intervalos sí contiene el valor real del parámetro.

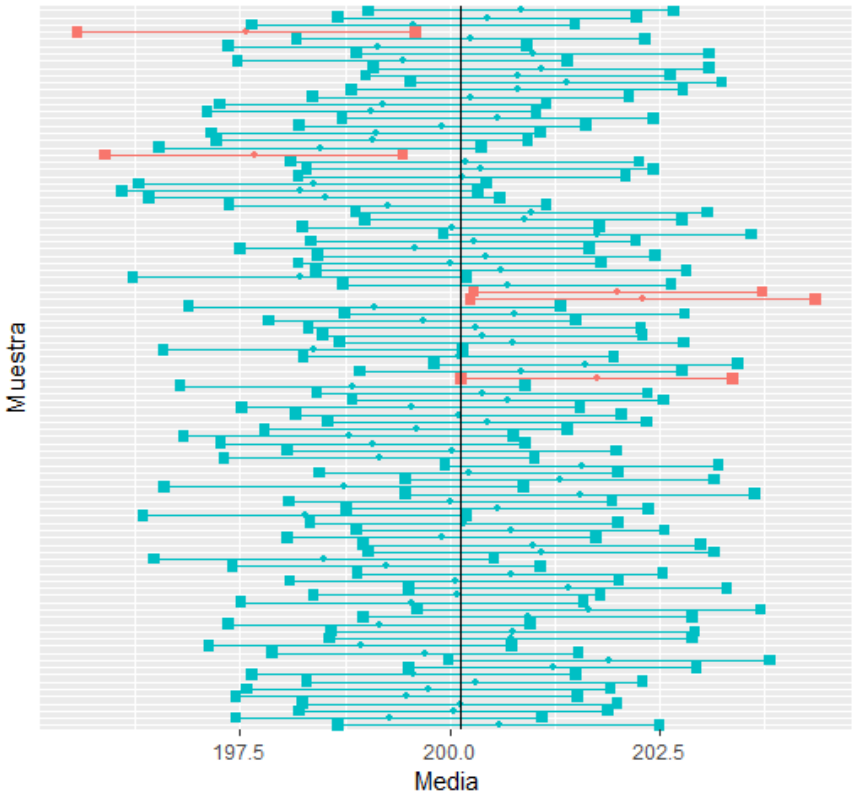

Figura 4.3: Intervalos de Confianza simulados

4.1.2 Intervalo de Confianza para la media μ

Para construir un intervalo de confianza para μ existen dos casos, el primero es cuando se conoce la desviación poblacional σ y el segundo cuando no se conoce la desviación poblacional σ.El primer caso es hipotético y puede ser considerado un caso para ejemplificar el concepto de intervalo de confianza, en la sección 4.1.2.1 se explica en detalle por qué consideramos este un caso hipotético. Para el primer caso usamos la distribución normal y para el segundo debemos usar otra distribución como lo veremos en la sección 4.1.2.2.

4.1.2.1 Intervalo de confianza para μ cuando se conoce σ

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n de una población con desviación conocida σ, un intervalo con $100\,(1-\alpha)\,\%$ para la media μ está dado por:

$$\left(\bar{x} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \tag{4.2}$$

Donde $Z_{\frac{\alpha}{2}}$ es el valor correspondiente a una probabilidad de cola superior de $\frac{\alpha}{2}$ de la distribución normal estándar. El valor de $Z_{\frac{\alpha}{2}}$ que se usa para construir un intervalo de confianza recibe el nombre de *valor crítico*.

Para utilizar la ecuación (4.2) es necesario conocer el valor de σ. Pero conocer σ implica conocer toodos los valores de la población. Y si se conocen todos los valores de la población se puede calcular el valor de la media poblacional que es lo que nos interesa estimar. En situaciones empresariales y financieras nunca se conoce la desviación estándar de la población y además las poblaciones son muy grandes lo que imposibilita poder examinar todos los valores. En la siguiente sección aprenderemos a abordar esta situación.

4.1.2.2 Intervalo de confianza para μ cuando no se conoce σ

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n con desviación muestral s, un intervalo con $100\,(1-\alpha)\,\%$ para la media μ está dado por:

$$\left(\bar{x} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) \tag{4.3}$$

donde $t_{\dfrac{\alpha}{2}}$ es el valor crítico correspondiente a una probabilidad de cola superior de $\dfrac{\alpha}{2}$ de la distribución t con $n-1$ grados de libertad.

4.1.3 Intervalo de Confianza para la proporción

Los dos intervalos de confianza vistos en las secciones anteriores son usados para variables cuantitativas, también se puede crear intervalos de confianza para variables categóricas. Por ejemplo, es posible que queramos estimar la proporción de elementos en una población que poseen cierta propiedad de interés. El parámetro de la proporción poblacional lo vamos a representar con la letra griega π. El estimador puntual para π es la proporción muestral $p = \frac{X}{n}$, donde n es el tamaño muestral y X es el número de elementos de la muestra que poseen la catacterística que interesa.

$$\left(p - Z_{\frac{\alpha}{2}}\sqrt{\frac{p\,(1-p)}{n}}, p + Z_{\frac{\alpha}{2}}\sqrt{\frac{p\,(1-p)}{n}}\right) \tag{4.4}$$

Donde

- $p =$ proporción muestral $= \dfrac{X}{n} = \dfrac{\text{Número de elementos con la característica}}{\text{Tamaño muestral}}$
- $n =$ tamaño muestral

4.1.4 Intervalo de Confianza para la diferencia de medias

Si tenemos dos poblaciones con media μ_1 y μ_2 y desviaciones σ_1 y σ_2 respectivamente un estimador puntual de la diferencia entre μ_1 y μ_2 viene dado por el estadístico $\bar{X}_1 - \bar{X}_2$. Es decir que para obtener un estimador puntual de $\mu_1 - \mu_2$ debemos escoger dos muestras independientes, una muestra de cada población, de tamaños n_1 y n_2.

De acuerdo al teorema del límite central $\bar{X}_1 - \bar{X}_2$ debe estar distribuida normalmente con media $\mu_1 - \mu_2$ y desviación $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

4.1.4.1 Desviaciones conocidas

Si $\bar{x_1}$ y $\bar{x_2}$ son medias de muestras aleatorias independientes de tamaños n_1 y n_2 de poblaciones con desviaciones conocidas σ_1 y σ_2, respectivamente, un intervalo de confianza al $100\,(1-\alpha)\,\%$ para $\mu_1 - \mu_2$ está dado por:

$$\left((\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \tag{4.5}$$

4.1.4.2 Desviaciones desconocidas e iguales

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes, de tamaños muestrales n_1 y n_2 respectivamente, de poblaciones aproximadamente normales con desviaciones desconocidas pero iguales un intervalo de confianza del $100\,(1-\alpha)\,\%$ para $\mu_1 - \mu_2$ está dado por

$$\left((\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \tag{4.6}$$

donde s_p es el estimador de la desviación conjunta y se calcula a partir de la expresión (4.7) y $t_{\frac{\alpha}{2}}$ es el valor t con $v = n_1 + n_2 - 2$ grados de libertad, con una probabilidad de $\frac{\alpha}{2}$. dejando un área de $\frac{\alpha}{2}$ a la derecha.

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} \tag{4.7}$$

4.1.4.3 Desviaciones desconocidas y diferentes

Si $\bar{x}_1$, s_1, $\bar{x}_2$ y s_2 son las medias y desviaciones de muestras aleatorias independientes de tamaños muestrales n_1 y n_2, respectivamente de poblaciones aproximadamente normales con varianzas desconocidas y diferentes un intervalo de confianza del $100\,(1-\alpha)\,\%$ para $\mu_1 - \mu_2$ está dado por:

$$\left((\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) \tag{4.8}$$

donde $t_{\frac{\alpha}{2}}$ es el valor t con

$$v = \left\lfloor \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \right\rfloor \tag{4.9}$$

grados de libertad, con un área de $\frac{\alpha}{2}$ a la derecha.

4.1.5 Intervalo de Confianza para la diferencia de proporciones

Si p_1 y p_2 son las proporciones de éxitos en muestras aleatorias de tamaño n_1 y n_2, respectivamente, $q_1 = 1 - p_1$, y $q_2 = 1 - p_2$ un intervalo de confianza del $100\,(1-\alpha)\,\%$ para la diferencia de dos proporciones poblacionales $\pi_1 - \pi_2$, está dado por

$$\left((p_1 - p_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}, (p_1 - p_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right) \tag{4.10}$$

4.2 Pruebas de hipótesis

Los intervalos de confianza pueden ayudar a contestar preguntas tales como "¿es razonable concluir que la media de los ingresos de las pequeñas empresas durante el año anterior es igual a medio millón de dólares?" y podrìamos verificar si el valor en cuestión está dentro del intervalo de confianza, de ser así podríamos contestar afirmativamente a la pregunta. Este procedimiento que acabamos de enunciar tiene mucho sentido pero puede resultar "muy débil" para contestar la pregunta. Preguntas como la mencionada anteriormente se pueden contestar usando pruebas de hipótesis. Una hipótesis, en términos estadísticos, es una suposición o afirmación sobre un parámetro de la población y una prueba de hipótesis es un procedimiento basado en evidencia de la muestra y la teoría de la probabilidad para determinar si la hipótesis es una afirmación razonable.

Toda prueba de hipótesis parte de un enunciado que se asume verdadero hasta probar lo contrario, esta hipótesis que se asume cierta es conocida como **hipótesis nula**. El nombre de *nula* proviene de la esperanza de que no exista diferencia significativa entre los grupos de prueba. La **hipótesis alternativa** es lo opuesto a la hipótesis nula. La hipótesis alternativa en la mayoría de los casos se condidera como una hipótesis de investigación.

La hipótesis nula se denota con H_0 y la alternativa con H_1 aunque algunos autores utilizan la notación H_a. Al final de una prueba de hipótesis se acepta o se rechaza la hipótesis nula, pero debemos tener claro que el procedimiento para probar hipótesis incluye la probabilidad de una conclusión incorrecta. Es decir que el investigador debe comprender que el rechazo de una hipótesis implica que la evidencia proporcionada por la muestra es la que rechaza la hipótesis, dicho de otra forma rechazar una hipótesis significa que existe una pequeña probabilidad de obtener la información muestral observada cuando, en realidad la hipótesis es verdadera.

Existen dos tipos de errores posibles en un procedimiento de prueba de hipótesis: rechazar la hipótesis nula cuando esta es verdadera y no rechazarla cuando esta es falsa, en la 4.1 se resume la relación entre las decisiones y los errores.

Tabla 4.1: Error tipo I y II

	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión Correcta	*Error tipo II*
Rechazar H_0	*Error Tipo I*	Decisión Correcta

El error de tipo I también se conoce como falso positivo y el error de tipo II como falso negativo. La probabilidad de cometer un error de tipo I se denota con α y la probabilidad de cometer un error de tipo II se denota con β.

Existen algunas propiedades importantes que se deben conocer sobre los errores de tipo I y tipo II.

- Los errores de tipo I y de tipo II están relacionados. Si la probabilidad de uno aumenta la probabilidad del otro disminuye.
- La probabilidad de cometer un error de tipo I se puede reducir ajustando el o los valores críticos.
- Un incremento del tamaño muestral n, reducirá α y β.
- Si la hipótesis nula es falsa, β tiene un máximo cuando el verdadero valor de un parámetro se aproxima al valor hipotético. A mayor distancia entre el verdadero valor y el hipotético, el valor de β sera menor.

Si se establece una buena regla de decisión, tendríamos una alta oportunidad de tomar una decisión correcta. Desafortunadamente no es posible eliminar completamente la posibilidad de errores, pues como se dijo anteriormente reducir la probabilidad de cometer el error de un tipo implica aumentar el error de otro tipo.

Retomemos la pregunta planteada al inicio de esta sección, tomemos como hipótesis nula que la media es igual a 500000 y como alternativa que la media es diferente de 500000, por el momento no entraremos en detalles de como escoger las hipótesis; las hipótesis serán expresadas en miles de dólares:

$$\begin{cases} H_0 : \mu = 500 \\ H_1 : \mu \neq 500 \end{cases} \tag{4.11}$$

Una *regla de decisión* podría ser escoger un punto o puntos límites en el eje $x´$como en la figura 4.4. La gráfica de color rojo representa H_0 y las de color celeste representan H_1. En este caso vamos a escoger dos puntos límites porque si la media muestral se encuentra a la izquierda o a la derecha de estos límites, se rechaza H_0. En cualquiera de los dos casos la media muestral está demasiado lejos de H_0 para ser creíble. Si la media muestral $\bar{x}$ está entre los valores límite se acepta H_0. La región entre los puntos límite es la *región de no rechazo* y hacía la izquierda o derecha de los puntos límite se encuentra la *región de rechazo.*

Partiendo de lo anterior, se pueden ver los errores de tipo I y II. El área de color verde bajo la distribución de H_0 muestra la probabilidad de rechazar H_0 cuando en realidad es verdadera, es decir un error de tipo I. El área de color celeste bajo la distribución de H_1 muestra la probabilidad de no rechazar H_0 cuando en realidad es falsa. En una prueba de hipótesis el investigador decide la probabilidad α de un error de tipo I, este valor de α es conocido como la **significancia**.

Los puntos lìmites son conocidos como **valores críticos**. Un procedimiento muy usado en pruebas de hipótesis es calcular el valor z de la media muestral, conocido como **estadìstico de prueba** y luego comparar el valor crìtico con este estadístico. Podrìamos resumir el procedimiento de prueba de hipótesis en 5 pasos:

1. Determinar las hipótesis nula y alternativa.
2. Escoger la significancia, generalmente se escoge el 5% o 0.05.
3. Determinar el estadístico
4. Calcular el valor crítico y establecer la regla de decisión
5. Tomar la decisión.

Una forma sencilla de entender cómo establecer la regla de decisión es ver la hipótesis alternativa. La hipótesis alternativa puede tener los símbolos $<$, $>$ o $\neq$. La región de rechazo de cada una se sombrea de color verde en las figuras 4.5, 4.6 y 4.7 respectivamente, al igual que en la figura 4.4 la distribución de rojo corresponde

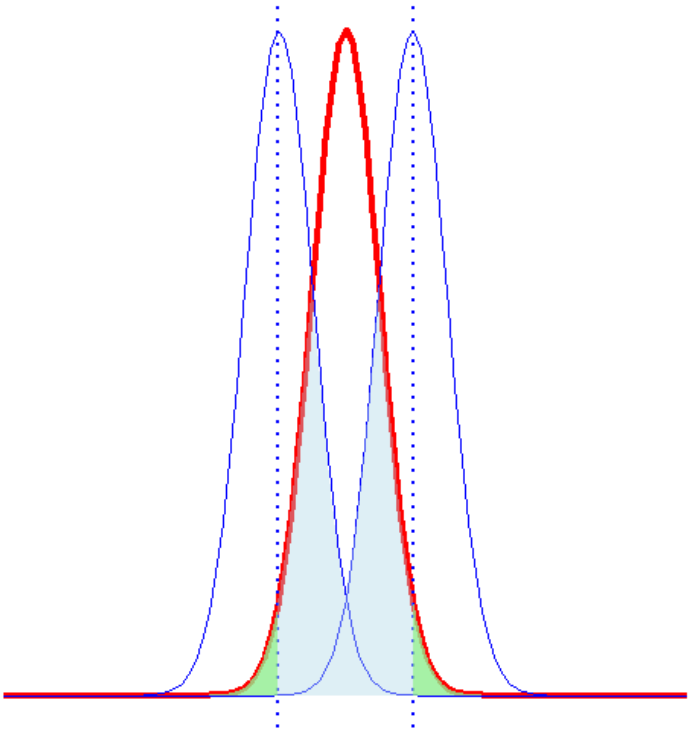

Figura 4.4: Errores Tipo I y II

a H_0 y la de azul a H_1. Las figuras 4.5 y 4.6 corresponden a pruebas de hipótesis de **una cola** y la figura 4.7 corresponde a una prueba de hipótesis de **dos colas**

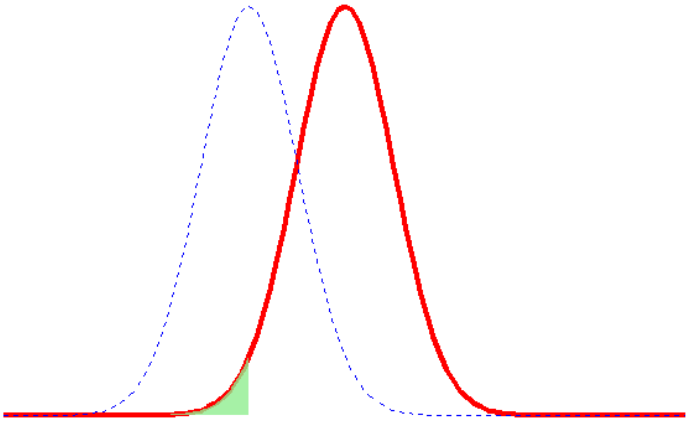

Figura 4.5: H_1 :<

4.2.1 Significancia, tamaño del efecto y potencia de la prueba

Los investigadores generalmente buscan resultados "significativos", es común encontrar en los artículos académicos, investigaciones y reportes expresiones como "los resultados son significativos", "el coeficiente es significativamente diferente de cero". Es importante entender que la palabra "significativo" se usa en términos estadisticos y no en el sentido de *importante* o *relevante*. Es decir que algo puede ser estadisticamente

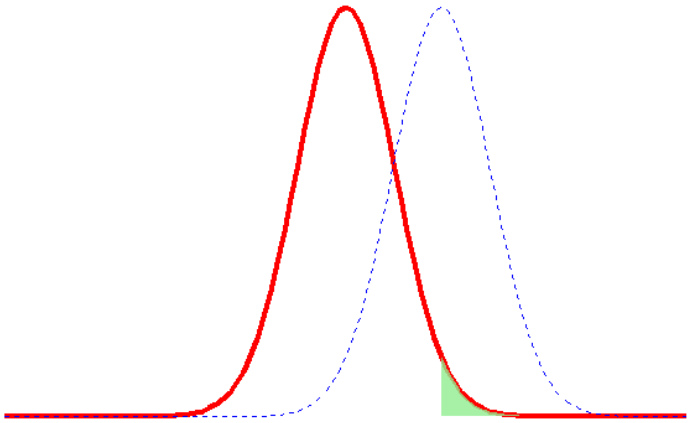

Figura 4.6: $H_1 :>$

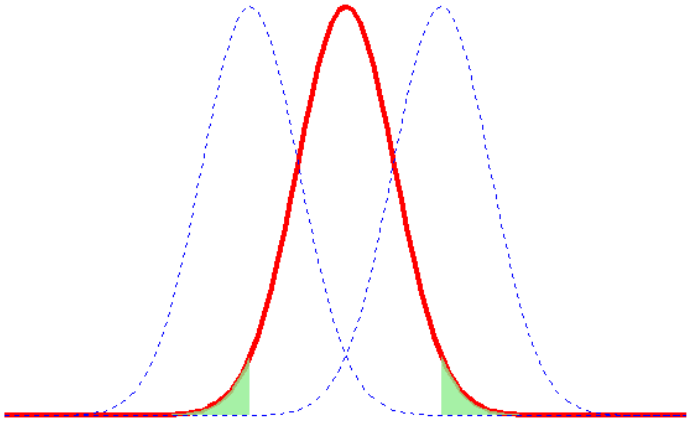

Figura 4.7: $H_1 :\neq$

significativo pero no relevante.

Supongamos que en el ejemplo desarrollado en la sección anterior tenemos una media muestral de 496 y que ese valor nos lleva a rechazar la hipótesis nula. Para nosotros esa diferencia puede no ser relevante, sin embargo esa diferencia es significativa en el sentido estadístico. Es decir que no solo deberíamos interpretar los resultados estadísticos solamente en términos de la significancia sino también considerar el tamaño de la diferencia esto es conocido como el **tamaño del efecto** y preguntarse si es importante o no.

Otro concepto relacionado a las probabilidades de error es la **potencia de una prueba** que es la probabilidad de rechazar H_0 cuando esta es falsa. Si la prueba fuera de una cola con hipótesis alternativa $>$, la potencia de la prueba viene dada entonces por el área bajo H_1 a la derecha de la línea de decisión.

$$\text{Potencia de una Prueba} = 1 - P\left(\text{Error Tipo II}\right) = 1 - \beta \tag{4.12}$$

4.2.2 El valor p

El valor p es la probabilidad, calculada suponiendo que la hipótesis nula es cierta, de obtener un valor del estadístico de prueba al menos tan contradictorio para H_0 como el valor calculado a partir de la muestra

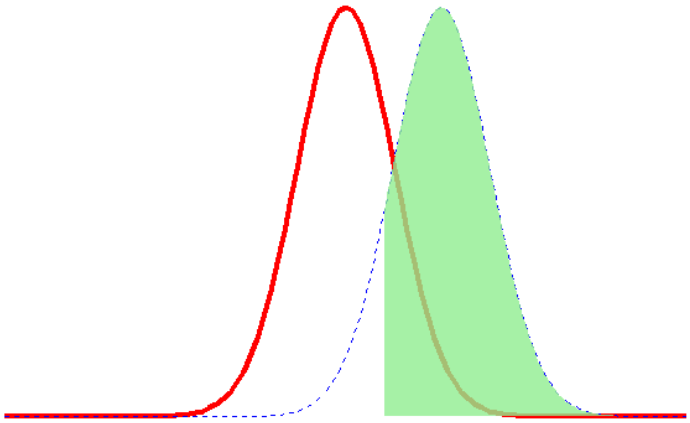

Figura 4.8: Potencia de una prueba de Hipótesis

disponible. Dicho de otra forma el valor p es el nivel más bajo de significancia en el que el valor observado del estadístico de prueba es significativo.

El criterio para rechazar H_0 utilizando el valor p es, si $p < \alpha$ se rechaza H_0. Es un criterio ampliamente usado, sin embargo en los últimos años ha tenido muchas críticas.

4.2.3 Estadísticos

Aunque el objetivo de este libro es trabajar con R, es necesario que el lector conozca los estadísticos usados en las diferentes pruebas de hipótesis que se trabajarán en las siguientes secciones,

Tabla 4.2: Estadísticos

Parámetro	Muestra	Varianza(S)	Estadístico	Distribución
Media	Grande	Conocida	$\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	Z
Media	Pequeña	Desconocidas	$\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$	t
Diferencia de Medias	Grande	Conocidas	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1}{n_1}+\frac{\sigma_2}{n_2}}}$	Z
Diferencia de Medias	Pequeña	Desconocidas e iguales	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	t
Diferencia de Medias	Pequeña	Desconocidas y diferentes	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	t
Proporción	No aplica	No aplica	$\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$	Z
Diferencia de proporciones	No apica	No aplica	$\frac{p_1-p_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	Z

4.3 Intervalos de confianza en R

4.3.1 Para la media

Recapitulando lo revisado en las secciones 4.1.2.1 y 4.1.2.2 se definieron los intervalos de confianza para la media con σ conocida y desconocida respectivamente. Como se mencionó antes el primer caso es un caso hipotético que sirve como una primera aproximación al concepto y la definición de intervalo de confianza. Mientras que el segundo caso es el escenario con el que frecuentemente se encuentra un investigador o un analista.

Para el primer caso `R` no tiene una función definida en su instalación base, a manera de ejemplo vamos a hacer algunas suposiciones para construir intervalos de confianza con la distribución normal. El paquete `BSDA` tiene la función `z-test( )` que también será utilizada. Mientras que para el segundo caso se tiene la función `t.test`, que como el lector ha de suponer tiene como objetivo principal realizar pruebas de hipóresis utilizando la distribución t de Student. Esta función además permite obtener intervalos de confianza.

En las dos subsecciones a continuación se trabajará con la variable `ACTIVOS` de las microempresas de la base de datos `Ranking2018Comercio.csv`

4.3.1.1 σ conocida

En la ecuación (4.2) se define el intervalo de confianza para la media cuando se conoce σ. Para construir el intervalo supondremos que la desviación de los datos es la desviación poblacional. Recalcamos que esta suposición la hacemos solo para efectos prácticos.

Primero cargamos los datos y seleccionamos solo los valores con los que deseamos trabajar. Usamos la función `filter( )` para seleccionar las observaciones que cumplen con la condición de que el tamaño de la empresa sea microempresa. Y luego seleccionamos solo la variable que analizaremos con la función `select( )`, en este caso solo escogemos la variable `ACTIVO`. Al finalizar la operación de filtro y selección usamos la función `attach( )`, el argumento que se escribe en la función es el nombre del conjunto de datos. La función `attach` sirve para cargar los datos a la memoria, la principal ventaja de usarla es que cuando los datos se cargan a la memoria no debemos usar el operador `$` por ejemplo en vez de escribir `dataframe$variable` ahora solo escribiremos `variable`.

```
rank18com = read.csv("Ranking2018Comercio.csv",header=TRUE,
                     dec=",", sep=";")

rank18com.Micro = rank18com %>%
  filter(TAMAÑO == "MICROEMPRESA") %>%
  select(ACTIVO)

attach(rank18com.Micro)
```

Vamos a calcular la media y la desviación de los activos de las microempresas.

```
media = mean(ACTIVO)
desviacion = sd(ACTIVO)
```

Vamos ahora a calcular el término $Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ para calcular el término $Z_{\frac{\alpha}{2}}$ trabajamos con la función `qnorm` el argumento que escribimos en la función es el valor $1-\frac{\alpha}{2}$, si el intervalo es con el 95% de confianza $\alpha = 0.05$ por lo que el valor a ser escrito en la función es $1-\frac{0.05}{2} = 0.975$. Para calcular $\sqrt{n}$ primero determinamos n con la función `nrow()` el argumento que recibe esta función es el conjunto de datos. Para calcular la raíz utilizamos la función `sqrt( )`.

```
error  = qnorm(0.975)*desviacion/sqrt(nrow(rank18com))
```

Ahora calculamos los extremos del intervalo:

```
menor = media - error
mayor = media + error
menor
## [1] 30831.37
mayor
## [1] 46981.15
```

El intervalo con el 95% de confianza para la media es $(29074.58, 48737.94)$. Si quisieramos calcular el intervalo con el 98% de confianza cambiamos el argumento en la función `qnorm()` de 0.975 a 0.99.

```
error  = qnorm(0.99)*desviacion/sqrt(nrow(rank18com))
menor = media - error
mayor = media + error
menor
## [1] 29321.9
mayor
## [1] 48490.62
```

El intervalo con el 98% de confianza para la media es $(27236.71, 50575.81)$.

Otra opción si no se desea programar todos los pasos es usar la función `z.test` del paquete `BSDA` en la función `z.test()` se debe indicar el valor de σ en este caso, como antes, asumiremos que σ es la desviación de los activos obtenida con anterioridad. Ya que solo nos interesa calcular el intervalo de confianza usaremos `z.test()$conf.int` que nos dará el intervalo con un 95% de confianza:

```
library(BSDA)
z.test(ACTIVO,sigma.x = desviacion)$conf.int
## [1] 29074.58 48737.94
## attr(,"conf.level")
## [1] 0.95
```

Si deseamos calcular el intervalo con un nivel de confianza diferente al 95%, por ejemplo el 98%, indicamos la opción `conf.level=0.98`

```
z.test(ACTIVO,sigma.x = desviacion, conf.level = 0.98)$conf.int
## [1] 27236.71 50575.81
## attr(,"conf.level")
## [1] 0.98
```

4.3.1.2 σ desconocida

La función `t.test()` se utiliza para realizar pruebas de hipótesis, sin embargo entre sus salidas o resultados se encuenta un intervalo de confianza, debido a que solo nos interesa calcular intervalos de confianza utilizaremos la instrucción `t.test()$conf.int` que nos dará directamente el intervalo con un 95% de confianza:

```
t.test(ACTIVO)$conf.int
## [1] 29039.57 48772.96
## attr(,"conf.level")
## [1] 0.95
```

Si deseamos calcular el intervalo con otro nivel de confianza, por ejemplo el 90%, agregamos dentro de la función `t.test()` la opción `conf.level = 0.90`

```
t.test(ACTIVO, conf.level = 0.90 )$conf.int
## [1] 30632.78 47179.74
## attr(,"conf.level")
## [1] 0.9
```

Para terminar esta sección vamos a liberar la memoria de los datos `rank18com.Micro` para esto usamos la función `detach`

```
detach(rank18com.Micro)
```

4.3.2 Diferencia de medias

4.3.2.1 σ_1 y σ_2 conocidas

Para esta sección necesitamos dos grupos el grupo 1 serán las pequeñas empresas y el grupo 2 las microempresas,

```
rank18com.Peq = rank18com %>%
  filter(TAMAÑO == "PEQUEÑA") %>%
  select(ACTIVO)
act.Peq = rank18com.Peq$ACTIVO

rank18com.Micro = rank18com %>%
  filter(TAMAÑO == "MICROEMPRESA") %>%
  select(ACTIVO)
act.Micro = rank18com.Micro$ACTIVO
```

Vamos a elaborar el intervalo de confianza para la diferencia de medias de los dos grupos $\mu_1 - \mu_2$. Esto lo vamos a hacer con la función `z.test` del paquete `BSDA`. En la sección 4.3.1.1 utilizamos la función para elaborar el intervalo de confianza. En extenso esta función se usa de la siguiente forma `z.test(x, y = NULL, alternative = "two.sided", mu = 0, sigma.x = NULL, sigma.y = NULL, conf.level = 0.95)` Como podemos ver por defecto el programa calcula el intervalo de confianza con un 95% de confianza. Además debemos indicar los valores de σ_1 y σ_2 y al igual que en la seccion 4.3.1.1 supondremos que la desviación muestral es la desviación poblacional. Vamos a calcular el intervalo de confianza para la diferencia de medias con el 95% de confianza.

```
library(BSDA)
z.test(x=act.Peq, sigma.x = sd(act.Peq),
       y=act.Micro, sigma.y = sd(act.Micro))$conf.int
```

```
## [1] 122739.5 204426.1
## attr(,"conf.level")
## [1] 0.95
```

Vamos ahora a modficar el parámetro de la confianza y trabajamos con un nivel de 90% de confianza.

```
library(BSDA)
z.test(x=act.Peq, sigma.x = sd(act.Peq),
       y=act.Micro, sigma.y = sd(act.Micro),
       conf.level = 0.90)$conf.int
```

```
## [1] 129306.0 197859.5
## attr(,"conf.level")
## [1] 0.9
```

4.3.2.2 Desviaciones desconocidas y diferentes

Para esta sección y la siguiente vamos a trabajar con las mismas variables de la sección anterior. Como en la sección 4.3.1.2 vamos a trabajar con la función `t.test` el uso de esta función es `t.test(x, y = NULL,alternative = c("two.sided", "less", "greater"),mu = 0, paired = FALSE, var.equal = FALSE,conf.level = 0.95, ...)` La función por defecto calcula el intervalo con el 95% de confianza y se asume que las desviaciones son diferentes.

Vamos a trabajar ahora con un intervalo de confianza al 95% para la diferencia de medias con desviaciones desconocidas y diferentes. Aunque por defecto la función asume que las desviaciones son diferentes incluimos la opción `var.equal = FALSE`.

```
t.test(x=act.Peq, y=act.Micro, var.equal = FALSE,
       conf.level = 0.95)$conf.int
```

```
## [1] 122396.1 204769.5
## attr(,"conf.level")
## [1] 0.95
```

Si calculamos el intervalo ahora con el 98% de confianza.

```
t.test(x=act.Peq, y=act.Micro, var.equal = FALSE,
       conf.level = 0.98)$conf.int
```

```
## [1] 114563.4 212602.2
## attr(,"conf.level")
## [1] 0.98
```

4.3.2.3 Desviaciones desconocidas e iguales

Cuando se asumen desviaciones iguales se usa la opción `var.equal = TRUE`. El intervalo de confianza con el 95% se calcula de la siguiente forma:

```
t.test(x=act.Peq, y=act.Micro, var.equal = TRUE,
       conf.level = 0.95)$conf.int
```

```
## [1] 134354.2 192811.4
## attr(,"conf.level")
## [1] 0.95
```

Mientras que el intervalo con el 98% seria:

```
t.test(x=act.Peq, y=act.Micro, var.equal = TRUE,
       conf.level = 0.98)$conf.int
```

```
## [1] 128861.3 198304.2
## attr(,"conf.level")
## [1] 0.98
```

4.4 Pruebas de Hipótesis en R

4.4.1 Para la media

En este caso vamos a usar la función `z.test()` o `t.test()` dependiendo si la o las desviaciones son conocidas o desconocidas respectivamente.

4.4.1.1 σ conocida

El uso de la función `z.test` es `z.test(x, y = NULL,alternative = "two.sided", mu = 0, sigma.x = NULL,sigma.y = NULL, conf.level = 0.95)` para pruebas de hipótesis de una sola variable solo ingresamos la variable a ser analizada en `x`, para indicar la hipótesis alternativa en la opción `alternative` podemos escribir `"greater"`para mayor, `"less"` para menor o `"two.sided"` para diferente. Como antes debemos indicar el valor de la desviación conocida que asumiremos también que es la desviación de la muestra. Para indicar el valor hipotético de la media reemplazamos en la opción `mu=0` el valor de 0 por el valor hipotético.

Para ejemplificar el uso de la función, supongamos que queremos verificar si la media del valor de los activos de las microempresas dedicadas al comercio es diferente de 40 000, con una significancia del 5%, cuando hacemos la prueba de hipótesis debemos indicar la confianza que es el complemento de la significancia Confianza $= 1 -$ Significancia $\rightarrow$ Significancia $= 1 -$ Confianza.

Como ya es costumbre, cargamos primero el archivo a ser analizado.

```
rank18com = read.csv("Ranking2018Comercio.csv",header=TRUE,
                     dec=",", sep=";")

rank18com.Micro = rank18com %>%
  filter(TAMAÑO == "MICROEMPRESA") %>%
  select(ACTIVO)

attach(rank18com.Micro)
```

La hipótesis nula y alternativa serían:

$$\begin{cases} H_0 : \mu = 40\,000 \\ H_1 : \mu \neq 40\,000 \end{cases} \tag{4.13}$$

La instrucción en R para hacer la prueba de hipótesis sería:

```
desviacion = sd(ACTIVO)
z.test(ACTIVO, mu = 40000, sigma.x = desviacion, conf.level = 0.95)
##
##  One-sample z-Test
##
## data:  ACTIVO
## z = -0.21804, p-value = 0.8274
## alternative hypothesis: true mean is not equal to 40000
## 95 percent confidence interval:
##  29074.58 48737.94
## sample estimates:
## mean of x
##  38906.26
```

En este caso el valor p (`p-value`) es igual a 0.8274 y recordemos que en la sección 4.2.2 se definió el criterio del valor p que consiste en rechazar H_0 cuando $p < \alpha$. En este caso se acepta la hipótesis nula de que el valor real de la media es igual a 40 000.

Supongamos ahora que se desea probar que el valor real de la media es menor a 40 000, para hacer esto incluimos la opción 'alternative =“less”.

```
z.test(ACTIVO, mu = 40000, sigma.x = desviacion,alternative = "less",  conf.level = 0.95)
##
##  One-sample z-Test
##
```

```
## data:  ACTIVO
## z = -0.21804, p-value = 0.4137
## alternative hypothesis: true mean is less than 40000
## 95 percent confidence interval:
##        NA 47157.26
## sample estimates:
## mean of x
##  38906.26
```

Las conclusiones son parecidas a la primera prueba de hipótesis realizada.

Capítulo 5

Correlación y Regresión

Supongamos que se quisiera analizar la relación entre las ventas y la utilidad de una empresa, para tratar de explicar de qué forma las ventas afectan la utilidad de las empresas. La existencia de la relación entre variables puede ser analizada usando un gráfico, una medida numérica o una ecuación. El gráfico utilizado para analizar la relación entre dos variables se llama *gráfico de dispersión* o *diagrama de dispersión*, en la figura 5.1 se muestran algunos diagramas de dispersión.

Un diagrama de dispersión muestra la relación entre dos variables cuantitativas medidas para las mismas observaciones. Los valores de una variable aparecen en el eje horizontal, y los valores de la otra variable aparecen en el eje verticcal. Cada observación aparece como el punto en el gráfico fijado por los valores de ambas variables para esa observación.

En un diagrama de dispersión podemos encontrar tres aspectos de la asociación entre dos variables.

1. La dirección de la relación. La relación entre dos variables puede no existir o ser negativa o positiva. Retomemos la figura 5.1 para entender la dirección de la relación. Cuando los valores de x e y aumentan al mismo tiempo se dice que es positiva las gráficas con los puntos de color azul muestran una relación positiva. Cuando los valores de x aumentan y los de y disminuyen, o viceversa se dice que la relación es negativa, las gráficas con los puntos de color rojo presentan una relación negativa. Cuando no existen patrones claros como en la gráfica donde los puntos se presentan de color negro, se dice que no existe relación.

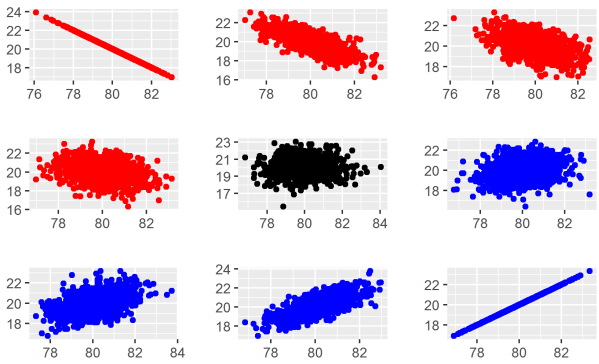

Figura 5.1: Gráficos de dispersión

2. La forma de la relación que puede ser lineal o no lineal. En la figura 5.2 se observa que la relación es lineal, mientras que en la figura 5.3 se observa una relación no lineal.

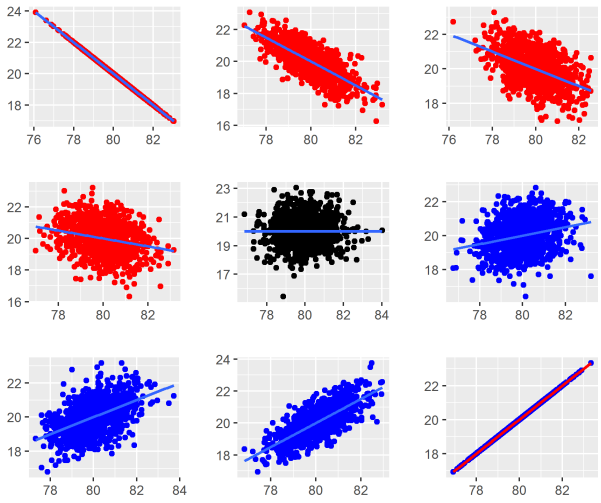

Figura 5.2: Relación Lineal

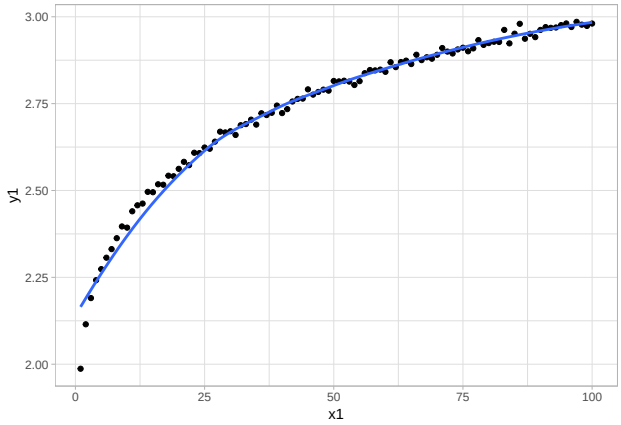

Figura 5.3: Relación No Lineal

3. La fuerza de la relación lineal. Imaginemos que trazamos una recta por el centro de la nube de puntos, la fuerza de la relación se puede medir por la proximidad de los datos a esa linea, a mayor cercanía a la recta mayor fuerza de la relación. En la figura 5.2 se observa que en las dos primeras y en las dos últimas gráficas las nubes de puntos se acercan hacia la recta, lo que nos indica que en esos casos la fuerza de la relación lineal es alta.

En los dos gráficos centrales dn la figura 5.4 la nube de puntos está más dispersa que en los otros gráficos la dispersión de la nube de puntos es menor. En términos de la fuerza de la relación lineal podemos decir que en los dos gráficos centrales esta es muy débil. Pero ¿cuán fuerte o débil es la relación en cada gráfica? La gráfica por si sola no nos indica que tan fuerte o débil es la relación, se hace necesario entonces medir de alguna forma la fuerza de la relación lineal. La fuerza de la relación lineal se mide con la correlación.

5.1 Coeficiente de Correlación

El *coeficiente de correlación de Pearson* es una medida que puede tomar valores entre -1 y 1. Es igual a 1 cuando dos variables cuantitativas tienen una relación lineal perfecta positiva y cuando las variables tienen

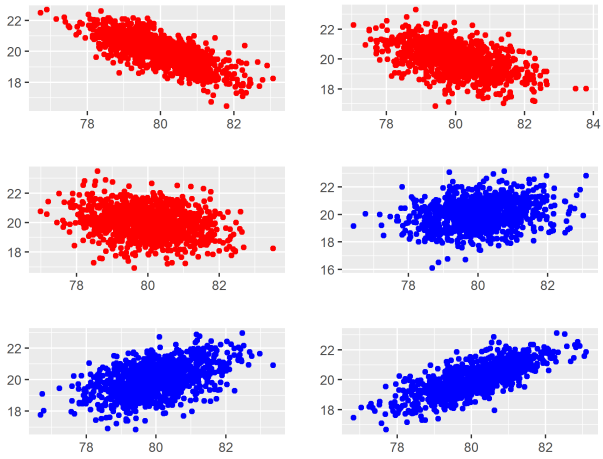

Figura 5.4: Fuerza de la relación lineal

una relación lineal negativa la correlación es igual a -1 como se observa en la figura 5.5.

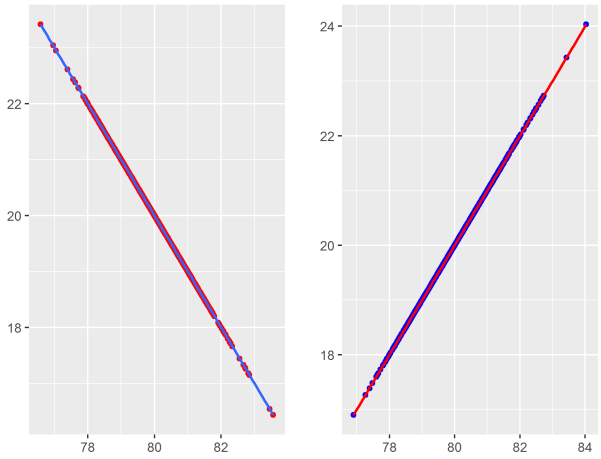

Figura 5.5: Relaciones lineales perfectas

Para entender el coeficiente de correlación de Pearson, debemos empezar por definir la *covarianza*. La varianza la definimos en la sección 2.7 como el promedio de la desviación cuadrática de todas las observaciones de una variable. Cuando trabajamos con dos variables debemos usar la **covarianza** que es la medida distancia entre cada par ordenado del centroide en un diagrama de dispersión. El centroide de un conjunto de puntos (x, y) es el punto $(\bar{x}, \bar{y})$.

En la figura 5.6 se presenta el diagrama de dispersión del gasto en Publicidad contra el ingreso por ventas de 100 empresas, estos datos han sido simulados. En la figura 5.7 se muestra la ubicación del centroide de los datos. Es fácil observar que el punto del centroide es el centro de un sistema de coordenadas con cuatro cuadrantes, como se muestra en la figura 5.7. La desviación de cualquier punto desde el centroide se calcula con la expresión $(x - \bar{x})(y - \bar{y})$.

Todos los puntos en el cuadrante I tienen publicidad y ventas mayores al promedio y cuando se reemplazan en la expresión $(x - \bar{x})(y - \bar{y})$ siempre se obtiene un resultado positivo. Los puntos en el cuadrante II tienen publicidad menor al promedio y ventas mayores al promedio al reemplazarlos en la expresión $(x - \bar{x})(y - \bar{y})$ se obtiene un resultado negativo puesto que negativo multiplicado por positivo es negativo, para los cuadrantes

III y IV los resultados de reemplazar en $(x - \bar{x})(y - \bar{y})$ son positivo y negativo respectivamente (¿por que?).

Para comparar las ventas con la publicidad se debe comparar la suma de los positivos obtenidos en los cuadrantes I y III con los negativos que resultan de los cuadrantes II y IV, Si la suma de los positivos es mayor a la de los negativos estamos ante una asociación positiva, pero si la suma de los negativos es mayor que la de los positivos la asociación es negativa y si la suma de los positivos es casi igual a la suma de los positivos, la sumatoria será cercana a 0 por lo que no existe asociación entre las variables.

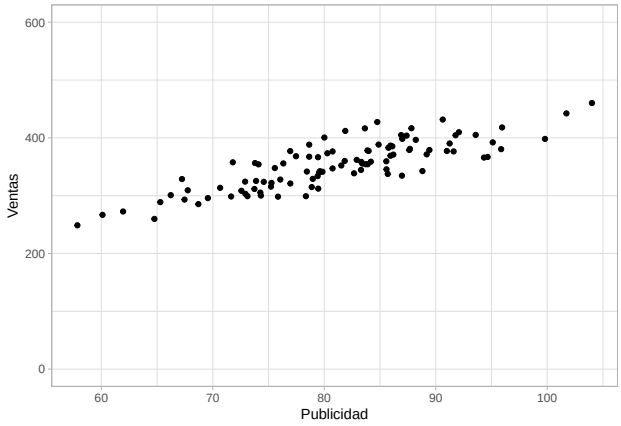

Figura 5.6: Diagrama de Dispersión del gasto en Publicidad contra el ingreso en Ventas

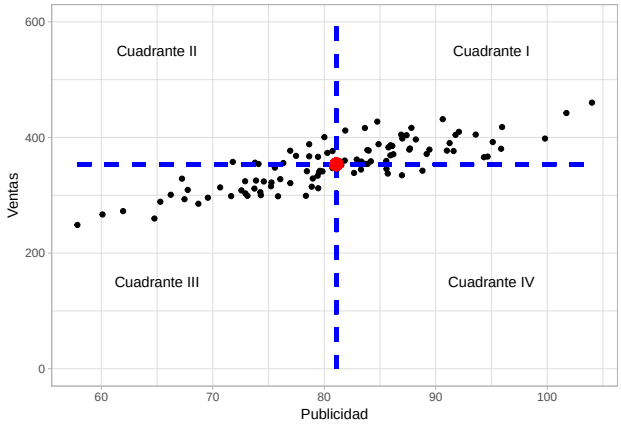

Figura 5.7: Ubicación del centroide de los datos

Es decir que la suma de las distancias entre los puntos y el centroide proporciona una medida de la relación entre las variables y si se divide este valor para el número de observaciones obtenemos la desviación promedio de los datos respecto al centroide, conocido como covarianza. Formalmente la covarianza se calcula con:

$$cov(x, y) = S_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{5.1}$$

La ecuación (5.1) puede ser reescrita como

$$S_{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \tag{5.2}$$

Si la covarianza es positiva, la relación entre las variables es positiva. Si la covarianza es negativa, la relación entre las variables puede ser negativa. Y si es 0 o cercana a 0 entonces no hay relación lineal entre las variables. Es decir que basta con conocer el signo de la covarianza para saber el sentido de la relación. Un problema de la covarianza es que su valor depende de las unidades de medida que están siendo usadas. Una forma de corregir esto es dividiendo para las desviaciones de x y y. El resultado de dividir la covarianza para las desviaciones recibe el nombre de **coeficiente de correlación de Pearson**

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \tag{5.3}$$

En R se utiliza la función `cor()`, vamos a hacer un ejemplo con los datos de la figura 5.6

```
set.seed(1.8)
Publicidad <- rnorm(100, mean=80, sd=10)
Ventas <- 30 + rnorm(100,mean=4,sd=0.3)*Publicidad + rnorm(100,mean = 0,sd=1)

cor(Publicidad, Ventas)
```

```
## [1] 0.834188
```

El valor obtenido de la correlación es 0.8342 pero ¿cómo saber qué tan bueno o malo es este valor de correlación? Es ampliamente aceptada la interpretación de acuerdo al intervalo en el que cae el valor de la correlación:

- -1: La relación lineal negativa es perfecta
- $-1 < r \leq -0.70$ La relación lineal negativa es fuerte
- $-0.70 < r \leq -0.50$ La relación lineal negativa es moderada
- $-0.50 < r \leq -0.30$ La relación lineal negativa es débil
- $-0.30 < r < 0$ La relación lineal es casi inexistente
- 0 No existe relación lineal
- $0 < r < 0.30$ La relación lineal es casi inexistente
- $0.30 \leq r < 0.50$ La relación lineal positiva es débil
- $0.50 \leq r < 0.70$ La relación lineal positiva es moderada
- $0.70 \leq r < 1$ La relación lineal positiva es fuerte
- 1 La relación lineal positiva es perfecta

5.2 Regresión lineal

Al inicio de este capítulo usamos diagramas de dispersión o gráficos de dispersión para analizar la relación entre dos variables. Luego en la seccion 5.1 se analizó la relación entre la publicidad y las ventas de 100 empresas utilizando el coeficiente de correlación. La correlación no contesta preguntas como ¿influye la publicidad sobre las ventas? Si la publicidad aumenta ¿cuánto aumentarán las ventas? En esta sección utilizaremos el análisis de regresión para analizar la relación entre la publicidad y las ventas.

El análisis de regresión es una forma más avanzada que la correlación para analizar la relación entre variables. Las principales diferencias entre la correlación y la regresión son:

- La regresión puede investigar las relaciones entre dos o más variables.
- Se estima una relación de causalidad entre la o las variables explicativas y la variable dependiente.
- Se mide la influencia de cada variable explicativa sobre la variable dependiente.
- Se puede medir la significancia de cada variable explicativa.

Se espera que la publicidad explique las ventas. Es decir que la publicidad es la variable explicativa o independiente y se ubica sobre el eje de las X, y el nivel de ventas es una variable explicada o dependiente y

se ubica sobre el eje de las Y. El análisis de regresión describe esta relación causal ajustando una línea recta a los datos como se observa en la figura 5.8. Esta recta de regresión es creciente, lo que se relaciona con el valor del coeficiente de correlación previamente calculado cuyo signo es positivo, es decir que altos niveles de publicidad se asocian con altos niveles de ventas y viceversa.

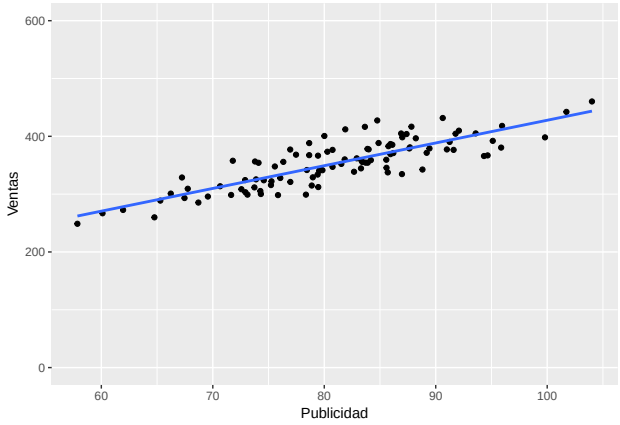

Figura 5.8: Recta de regresión

En los datos cuando la publicidad es 64.76 las ventas son 250.35, sin embargo para la recta de regresión cuando la publicidad es igual a 64.76 ventas son iguales a 289.29. Este valor es cercano pero no igual al valor real, la diferencia refleja la ausencia de una correlación perfecta entre las dos variables. La diferencia entre el valor real Y y el valor predicho $\hat{Y}$ recibe el nombre de **error** o **residual**, en la figura 5.9 se observa el error.

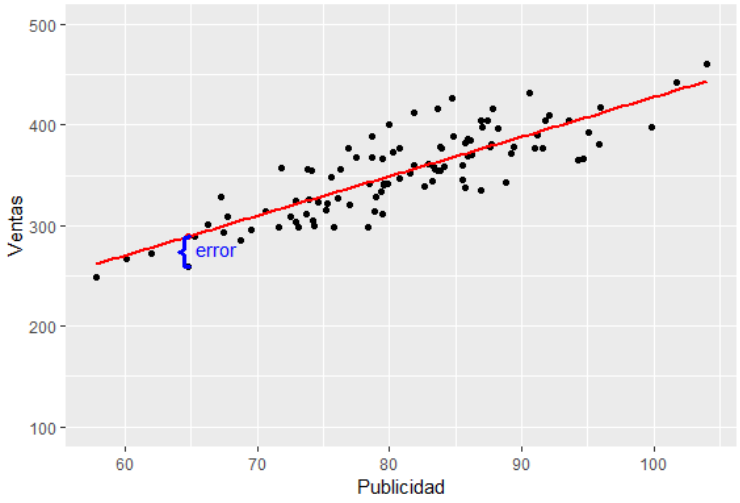

Figura 5.9: Error en la estimación

El error puede provenir de diversas fuentes y deberse a varios factores por ejemplo en el contexto que estamos analizando, la diferencia entre las ventas predichas y las ventas reales se puede deber a errores de medición, a condiciones externas a la empresa como la situación política o a condiciones internas como problemas en la producción, etc. Todos los factores están dentro del término de error y esto significa que las observaciones caen alrededor de la recta de regresión y no en ella. Si hay muchos de estos factores, sin uno en particular que predomine y además existe independencia entre los factores se puede asumir que los errores se distribuyen normalmente alrededor de la recta de regresión.

$$\hat{Y}_i = \beta_0 + \beta_1 X_i \tag{5.4}$$

donde:

- $\hat{Y}_i$ es el valor predicho de Y para la observación i
- X_i es el valor de la variable explicativa para la observación i
- β_0 y β_1 son los coeficientes fijos que serán estimados; β_0 representa el intercepto de la recta de regresión con el eje de las Y y β_1 mide la pendiente. En la figura 5.10 se aprecia el intercepto y la pendiente de la recta de regresión lineal.

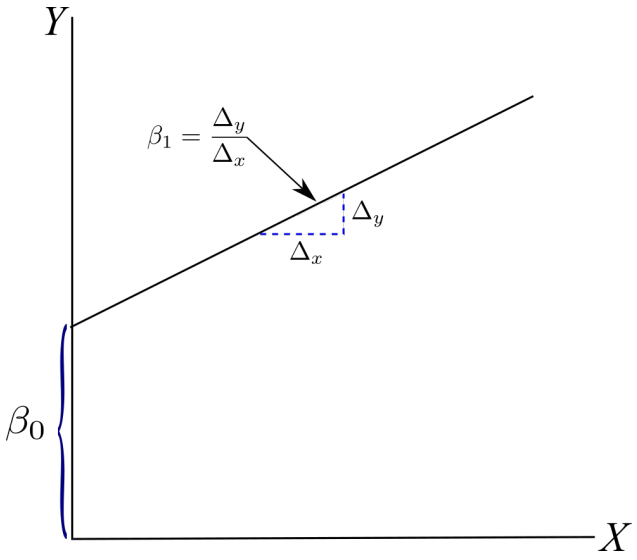

Figura 5.10: Intercepto y pendiente en la recta de regresión

Lo primordial del análisis de regresión consiste en determinar los valores de β_0 y β_1. Para esto se parte de que la diferencia entre el valor real y el valor predicho es igual al error o dicho de otra forma el valor real es igual al valor predicho más el error es decir:

$$Y_i = \hat{Y}_i + e_i \tag{5.5}$$

Reemplazando la ecuación (5.5) en (5.4) se obtiene:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \tag{5.6}$$

Con la ecuación (5.6) se puede entender que las ventas observadas están conformadas de dos componentes:

1. La parte que se explica por la publicidad $\beta_0 + \beta_1 X_i$
2. El error e_i

La recta de mejor ajuste se determina encontrando los valores β_0 y β_1 que **minimizan** la suma de los **errores cuadráticos** es decir $\sum_{i=1}^{n} e_i^2$, este método se lo conoce como **mínimos cuadrados ordinarios**. Intuitivamente la primera idea sería minimizar la suma de los errores $\sum_{i=1}^{n} e_i$ sin embargo $\sum_{i=1}^{n} e_i = 0$.

De la ecuación (5.6) se obtiene que

$$e_i = Y_i - \beta_0 - \beta_1 X_i \tag{5.7}$$

Entonces la suma de los errores cuadráticos equivale a:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{5.8}$$

La expresión de la ecuación (5.8) se minimiza utilizando derivadas, el resultado obtenido de esa minimización se muestra en las ecuaciones (5.9) y (5.10)

$$\beta_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \tag{5.9}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \tag{5.10}$$

5.2.1 Regresión linear en R

Para realizar regresión lineal en R se puede usar la función `lm( )`, el uso de la función es `lm(formula, datos)`. En este caso se quiere obtener los coeficientes de la ecuación de regresión

$$\text{Ventas} = \beta_0 + \beta_1 \text{Publicidad} + \epsilon \tag{5.11}$$

```
set.seed(1.8)
Publicidad <- rnorm(100, mean=80, sd=10)
Ventas <- 30 + rnorm(100,mean=4,sd=0.3)*Publicidad + rnorm(100,mean = 0,sd=1)

m1 = lm(Ventas ~ Publicidad)
summary(m1)
```

```
##
## Call:
## lm(formula = Ventas ~ Publicidad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.588 -15.371  -3.649  14.335  59.570
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.6898    21.4169    1.62    0.109
## Publicidad    3.9312     0.2625   14.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.46 on 98 degrees of freedom
## Multiple R-squared:  0.6959, Adjusted R-squared:  0.6928
## F-statistic: 224.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

5.2.2 Interpretación de los coeficientes de regresión

En el apartado `Coefficients:` del resultado la primera columna corresponde a los valores estimados de los coeficientes. La ecuación de regresión obtenida al reemplazar los coeficientes obtenidos en la ecuación (5.11) seria:

$$\text{Ventas} = 34.69 + 3.93 \text{Publicidad} + \epsilon \tag{5.12}$$

El primer coeficiente a interpretar es el de la pendiente en este caso β_1 es igual a 3.93 esto se lo observa en la salida de R en el apartado `Coefficients:` el valor estimado para la publicidad, de manera general

el valor de β_1 se interpreta como el cambio en la variable dependiente (Y) por cada aumento de 1 unidad en la variable independiente (X). Es decir que en este caso por cada aumento de 1 dólar en la inversión en publicidad, el nivel de ventas sube en promedio 3.93.

Por otro lado β_0 es igual a 34.69, β_0 se interpreta como el valor que toma la variable dependiente cuando la variable independiente es igual a 0. En este caso cuando la inversión en publicidad es igual a 0 las ventas son en promedio iguales a 34.69.

5.2.3 Bondad de Ajuste del modelo de regresión

Una vez calculada la recta de regresión quizás nos preguntemos si esta da un buen ajuste a los datos, es decir si las observaciones se alejan o se acercan de la recta. Si el ajuste es pobre, quizás el efecto de la variable independiente en la dependiente no es lo suficientemente fuerte. El lector debe fijarse en que, aún cuando no haya efecto de X sobre Y se puede calcular la recta de regresión. Medir la bondad de ajuste de una regresión nos permite discriminar entre un buen y un mal modelo de regresión.

La bondad de ajuste se calcula comparando dos rectas, la recta de regresión y la recta promedio de Y que es una recta horizontal dibujada en el promedio de Y, como se observa en la figura 5.11 la linea punteada de negro representa el valor promedio de Y, se ilustra además una observación (X_i, Y_i). La diferencia entre Y y $\bar{Y}$ está dividida en dos partes la primera parte es la parte que explica la recta de regresión $\hat{Y} - \bar{Y}$. y la segunda parte es el término de error $Y - \hat{Y}$

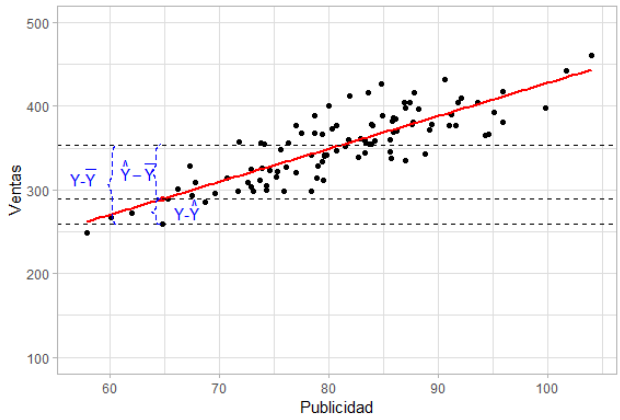

Figura 5.11: Bondad de Ajuste de un Modelo de Regresión Lineal

Un buen modelo de regresión debería explicar una gran porción de las diferencias entre Y y $\bar{Y}$, por lo tanto la longitud $\hat{Y} - \bar{Y}$ debería ser más grande respecto a $Y - \bar{Y}$. Entonces una medida de ajuste puede ser $\frac{\hat{Y}-\bar{Y}}{Y_i-\bar{Y}}$, esto lo deberíamos usar para todas las observaciones por lo que deberiamos sumar estas expresiones sin embargo en el caso de $\sum_{i=1}^{n} Y_i - \bar{Y}$ esto es igual a 0, para saltar este problema usamos el cuadrado de las expresiones para que se hagan positivas. De esta forma definimos:

1. $\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$, conocido como la suma total de cuadrados (STC).
2. $\sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2$, suma cuadrática del error (SCE)

La bondad de ajuste R^2 se define como:

$$R^2 = \frac{STC - SCE}{STC} = 1 - \frac{SCE}{STC} \tag{5.13}$$

El R^2 se lo interpreta como la proporción de la variabilidad de Y explicada por X. En el modelo obtenido en la sección 5.2.2 se obtuvo un R^2 de 0.6959 este valor se observa en las últimas líneas del resumen del modelo

donde dice `Multiple R-squared`, se interpreta que el 69.59% de la variabilidad de las ventas son explicadas por la inversión en publicidad.

5.2.4 Pruebas de Hipótesis para los coeficientes de la regresión

El resumen de R muestra los valores p para las hipótesis nulas $\beta_0 = 0$ y $\beta_1 = 0$, el criterio que se usa para aceptar o rechazar la hipótesis nula es el criterio que se explica en la sección 4.2.2.

En este caso se puede ver en el resumen que para el intercepto β_0 el valor p es 0.109 y para la pendiente β_1 el valor p es $< 2 \times 10^{-16}$ junto a los valores p hay unos códigos de significancia lo mejor es simplemnte usar el criterio ya mencionado, en este caso el valor p del intercepto nos indica que no se puede rechazar $H_0 : \beta_0 = 0$ a un nivel de significancia de 0.05, mientras que para el caso de la pendiente el valor p nos indica que rechazamos $H_0 : \beta_1 = 0$ a un nivel de significancia de 0.05.

Además en la última linea de la salida del resumen del modelo de regresión se obtiene un estadístico F de 224.2 este estadístico es un buen indicador de si existe una relación entre la variable independiente y dependiente. A mayor distancia del estadístico de 1 el modelo es mejor. Qué tan grande debe ser el valor depende tanto del número de datos y de variables predictoras. En este caso el estadístico está lejos de 1, además el valor p general al final del modelo es para el estadístico F en este caso el valor p de $< 2 \times 10^{-16}$ nos indica que el modelo global es significativo.

5.3 Regresión Múltiple

La regresión lineal simple sólo permite tener una variable explicativa, pero en la vida real una variable de respuesta puede ser afectada por más de una variable explicativa. Por ejemplo la demanda puede verse afectada por el precio, pero también por el ingreso. Es decir que la demanda puede ser expresada en función del precio y de los ingresos $Demanda = f\,(Precio, Ingreso)$. La ecuación de regresión puede ser ahora escrita como:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} \tag{5.14}$$

El subíndice k, i se entiende como el valor de la variable k para la i-ésima observación. Como en la regresión simple Y_i puede ser escrito como:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \epsilon_i \tag{5.15}$$

Combinando las ecuaciones (5.14) y (5.15). Se obtiene:

$$Y_i = \hat{Y}_i + \epsilon_i \tag{5.16}$$

Los principios usados en regresión múltiple son esencialmente los mismos que la regresión simple. Se determinan los coeficientes $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ minimizando la suma de los errores cuadráticos. También se obtiene un R^2 y se pueden realizar pruebas de hipótesis para los coeficientes.

Al igual que en la regresión lineal simple, la fracción de la variación en la variable dependiente explicada por las variables independientes se lo puede calcular como en la ecuación **??** es decir:

$$R^2 = 1 - \frac{SCE}{STC} \tag{5.17}$$

Si se añaden nuevas variables independientes a la regresión el valor del R^2 aumentará. Para entender esto, podemos pensar en un experimento en el cual añadimos variables independientes hasta que el total de variables independientes más la constante es igual al número de observaciones. Esto nos conducirá a un R^2 igual a 1. Entonces para poder explicar más de la variación de la variable dependiente simplemente adicionamos variables independientes sin importar si estas variables añadidas son realmente relevantes para la variable dependiente. Para obtener una medida más significativa de cuanta variación de la variable dependiente está siendo explicada, se ajusta el R^2 para compensar la pérdida de grados de libertad asociados a la inclusión de variables independientes adicionales. Este R^2 ajustado se calcula con la expresión de la ecuación (5.18)

$$R^2 = 1 - \frac{n-1}{n-k-1}\frac{SCE}{STC} \tag{5.18}$$

La regresión múltiple se hace también con la función `lm(formula,datos)`, la diferencia es que en la fórmula se incluye más de una variable independiente.

Para ejemplificar la regresión múltiple en R vamos a construir un modelo de regresión múltiple con la base `Ranking2018Comercio.csv`, se intentará explicar primero la utilidad de las empresas en función del número de empleados y las ventas. Es decir que la ecuación de regresión sería:

$$\text{Utilidad} = \beta_0 + \beta_1 \text{Empleados} + \beta_2 \text{Ventas} + \epsilon \tag{5.19}$$

Primero cargamos los datos. Una particularidad que tienen los datos es que algunas de las observaciones son inconsistentes en el sentido que reportan ventas iguales a 0, se filtrarán los datos de tal forma que solo se analizarán las empresas que han reportado ventas mayores a 0. Se seleccionará además la variable `TAMAÑO` para verificar más adelante si el tamaño de la empresa influye sobre las utilidades:

```
datos = read.csv("Ranking2018Comercio.csv",header=TRUE,sep=";", dec=",")

datos = datos %>%
  filter(VENTAS>0) %>%
  select(UTILIDAD,EMPLEADOS,VENTAS,TAMAÑO)
attach(datos)
```

Una vez cargados los datos vamos a graficar diagramas de dispersión por pares de variables para esto usaremos la función `pairs()` las variables para las que se quiere realizar son las tres primeras por lo que las seleccionamos indicando que del conjunto datos solo queremos las columnas de la 1 a la 3. En la figura 5.12 se observa que cuando se relaciona la utilidad con los empleados si hay una tendencia creciente bien definida, sin embargo cuando se relacionan las ventas ya sea con utilidad o con empleados la tendencia no está bien definida.

```
pairs(datos[,1:3])
```

Ahora construimos el modelo de regresión. Nótese que en la fórmula del lado de las variables independientes incluimos las dos variables ya mencionadas:

```
m2 = lm(UTILIDAD ~ EMPLEADOS + VENTAS)
summary(m2)
```

```
##
## Call:
## lm(formula = UTILIDAD ~ EMPLEADOS + VENTAS)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1193229     1031    12491    18190   484129
##
```

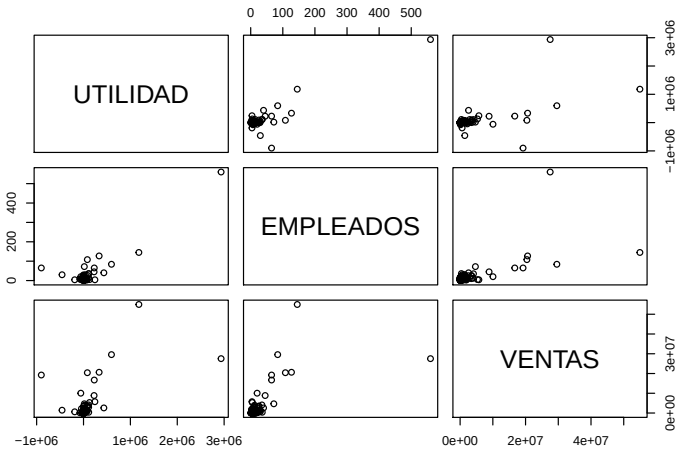

Figura 5.12: Diagramas de Dispersión del Conjunto datos

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.209e+04  6.344e+03  -5.059 7.71e-07 ***
## EMPLEADOS    4.900e+03  2.203e+02  22.237  < 2e-16 ***
## VENTAS       3.343e-04  1.709e-03   0.196    0.845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101100 on 276 degrees of freedom
## Multiple R-squared:  0.7622, Adjusted R-squared:  0.7605
## F-statistic: 442.4 on 2 and 276 DF,  p-value: < 2.2e-16
```

Se puede apreciar que con los coeficientes obtenidos la ecuación (5.19) sería:

$$\text{Utilidad} = -32\,090 + 4\,900\text{Empleados} + 0.0003343\text{Ventas} + \epsilon \tag{5.20}$$

El intercepto y el coeficiente estimado para el número de empleados son significativos, no así el estimado para las ventas. La interpretación de cada coeficiente estimado es:

- Intercepto: una empresa sin empleados y con ventas iguales a 0 tiene utilidades de $-32\,090$ (Pérdidas)
- Empleados: por cada empleado adicional que tiene la empresa las utilidades aumentan en promedio 4 900 dólares
- Ventas: por cada dólar adicional de ventas las utilidades aumentan en promedio 0.0003343

El R^2 ajustado del modelo es 0.7605 es decir que el 76.05% de la variación de la utilidad es explicada por el número de empleados y las ventas. De acuerdo al estadístico F y al valor p de ese estadístico el modelo global es significativo.

Capítulo 6

Análisis Factorial

6.1 Análisis de Fiabilidad

6.2 Evaluación de Análisis Factorial

Capítulo 7

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.