# Resumen sobre estadística y probabilidad.

Leandro Molina, Martin Borgo

25 de mayo de 2023


[1]

[1] Personaje perteneciente a Twin-Sensei

# Índice

# 1. Capitulo 1. ¿Que es la estadística?

## 1.1. ¿Por qué se debe estudiar estadística?

Hay 3 motivos para el estudio de la estadística estos son:

1. La primera razón consiste en que la información numérica prolifera por todas partes. si revisas diarios o revistas contienen mucha cantidad de información numérica.
2. Una segunda razón, es que las técnicas de la estadística se emplean para tomar decisiones que afectan la vida diaria, es decir, que incluyen en su bienestar.
3. Una tercera razón, el conocimiento de sus métodos facilita la compresión de la forma en que se toman las decisiones y proporciona un entendimiento mas claro de como le afectan.

Al encarar la necesidad de tomar decisiones en las que tenes que saber hacer un análisis de datos resultara de utilidad. Con el fin de tomar una decisión informada, sera necesario llevar a cabo lo siguiente para poder tomar una decisión informada:

1. Determinar si existe información adecuada o si requiere información adicional.
2. Reunir información adicional, si se necesita, de manera que no se obtengan resultados erróneos.
3. Resumir los datos de manera útil e informativa.
4. Analizar la información disponible.
5. Obtener conclusiones y hacer inferencias al mismo tiempo que se evaluá el riesgo de tomar una decisión incorrecta.

En resumen hay por lo menos tres razones para estudiar estadística: 1) los datos proliferan por todas partes; 2) las técnicas estadísticas se emplean en la toma de decisiones que influyen en su vida; 3) sin que importe la carrera que elija, tomara decisiones profesionales que incluyan datos.

## 1.2. ¿Que se entiende por estadística?

Posee dos significados: su aceptación más común, la estadística se refiere a información numérica. Una colección de información numérica recibe el nombre de **estadísticas**. La información estadística se presenta en forma gráfica, es útil porque capta la atención del lector e incluye una gran cantidad de información.

**Estadística:** Ciencia que recoge, organiza, presenta, analiza e interpreta datos con el fin de propiciar una toma de decisiones mas eficaz.

El primer paso en el estudio de un problema consiste en recoger datos relevantes. Estos deben organizarse de alguna forma y, tal vez, representarse en una gráfica.

## 1.3. Tipos de estadística.

El estudio de la estadística se divide en dos categorías: la estadística descriptiva y la estadística inferencial.

**Estadística descriptiva.**

Es la ciencia que recoge, organiza, presenta, analiza...datos". Esta parte de la estadística recibe el nombre de **estadística descriptiva**.

**Estadística descriptiva:** Métodos para organizar, resumir y presentar datos de manera informativa.

Se trata de estadística descriptiva si calcula el crecimiento porcentual de una década a otra. Sin embargo, no seria de naturaleza descriptiva si utiliza estos para el calcular con esos datos algo futuro. Una masa de datos desorganizados resulta de poca utilidad. Las técnicas de la estadística descriptiva permiten organizar esta clase de datos y darles significado. Los datos se ordenan en una **distribución de frecuencia** (mas adelante lo veremos). Se emplean diversas clases de **graficas** para describir datos.

**Estadística inferencial.**

La estadística inferencial, también denominada **inferencia estadística**. El principal interés que despierta esta disciplina se relaciona con encontrar algo relacionado con una población a partir de una muestra de ella. Ya que estas son inferencias relacionadas con una población, basadas en datos de la muestra, se trata de estadística inferencial. Se podría considerar a la estadística inferencial como la mejor conjetura que es posible obtener del valor de una población sobre la base de la información de una muestra.

**Estadística inferencial:** Métodos que se emplean para determinar una propiedad de una **población** con base en la información de una **muestra** de ella.

Atención a las palabras población y muestra en la definición de estadística inferencial. Una **población** puede constar de individuos, también puede consistir en objetos. Desde una perspectiva estadística, una población no siempre que tiene que ver con personas.

**Población:** Conjunto de individuos u objetos de interés o medidas que se obtienen a partir de todos los medios u objetos de interés.

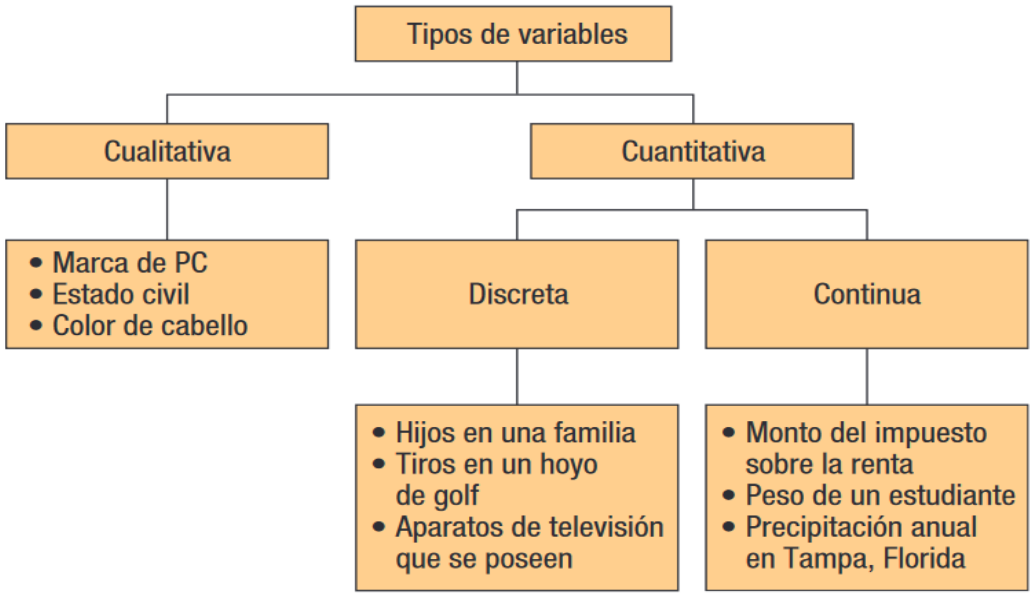
Con el objeto de inferir algo sobre una población, lo común es que se tome una muestra de ella.

**Muestra:** Porción o parte de la población de interés.

La toma de muestras para aprender algo sobre una población es de uso frecuente en administración, agricultura, política y acciones de gobierno.

## 1.4. Tipos de variables.

Una variable es una característica observable en las unidades estadísticas y tiene, por lo menos, dos valores.
Dos tipos básicos de variables: 1)Cualitativas y 2)Cuantitativas, la característica que se estudia es de naturaleza no numérica, recibe el nombre de **variable cualitativa** o **atributo**. Cuando los datos son de naturaleza cualitativa, importa la cantidad o proporción que caen dentro de cada categoría. Los datos cualitativos se resumen en tablas o graficas de barras. Cuando la variable que se estudia aparece en forma numérica, se le denomina **variable cuantitativa**. Las variables cuantitativas pueden ser discretas o continuas. Las **variables discretas** adoptan solo ciertos valores y existen vacíos entre ellos. Las variables discretas son el resultados de una relación numérica, las observaciones de una **variable continua** toman cualquier valor dentro de un intervalo especifico. Por lo general las variables continuas son el resultado de mediciones.
Resumen de los tipos de variables:




## 1.5. Niveles de medición.

Los datos se clasifican por niveles de medición. El nivel de medición de los datos rige los cálculos que se llevan a cabo con el fin de resumir y presentar los datos. También determina las pruebas estadísticas que se deben realizar. De hecho, existen cuatro niveles de medición: nominal, ordinal, de intervalo y de razón. La medición mas baja, o mas primaria, corresponde al nivel nominal. La mas alta, o el nivel que proporciona la mayor información relacionada con la observación, es la medición de razón.

**Datos de nivel nominal.**

Las observaciones acerca de una variable cualitativa solo se clasifican y se cuentan. No existe una forma particular para ordenar las etiquetas, no existe un orden natural. Para el nivel nominal, la medición consiste en contar, a veces, para una mejor compresión de lectura, estos conteos se convierten en porcentajes. Es necesario hacer que el porcentaje sume un total de 100 %, no existe un orden natural para los resultados. Para procesar datos a menudo se codifica la información en forma numérica. El nivel nominal tiene las siguientes propiedades:

1. La variable de interés se divide en categorías o resultados.
2. No existe un orden natural de los resultados.
3. Los datos solo se clasifican.

**Datos de nivel ordinal.**

El nivel inmediato superior de datos es el **nivel ordinal**. No es posible distinguir la magnitud de las diferencias entre los grupos, ¿la diferencia entre superior y bueno es la misma que entre lo malo e inferior? No es posible afirmarlo. Las propiedades del nivel ordinal de los datos son las siguientes:

1. Las clasificaciones de los datos se encuentran representadas por conjuntos de etiquetas o nombre (alto, medio, bajo), las cuales tienen valores relativos.
2. En consecuencia, los valores relativos de los datos se pueden clasificar u ordenar.

**Datos de nivel de intervalo.**

**El nivel de intervalo** de medición es el nivel inmediato superior. Incluye todas las características de nivel ordinar, pero, ademas, la diferencia entre valores constituye una magnitud constante. Si las distancias entre los números tienen sentido, aunque las razones no, entonces tiene una escala de intervalo de medición. Las propiedades de los datos de nivel intervalo son las siguientes:

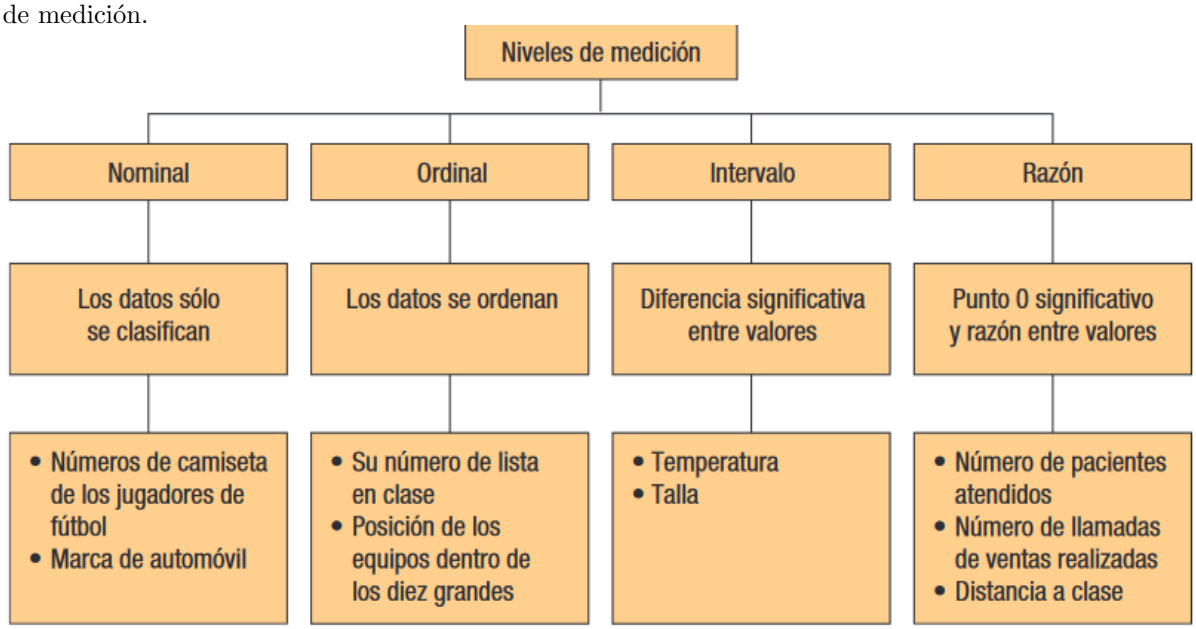
1. Las clasificaciones de datos se ordenan de acuerdo con el grado que posea de las característica en cuestión.
2. Diferencias iguales en la característica representan diferencias iguales en las mediciones (es decir la diferencia entre valores es significativa).
3. El cero es relativo (no indica ausencia de estado).

**Datos de nivel de razón.**

Todos los datos cuantitativos son registrados en el nivel de razón de la medición, el **nivel de razón** es el *mas alto*. Posee todas las características del nivel de intervalo, aunque, ademas, el punto 0 tiene sentido y la razón entre dos números significativa, si se encuentra en 0 significa la ausencia de algo (peso, dinero, etc). Las propiedades de los datos de nivel intervalo son las siguientes:

1. Las clasificaciones de datos se ordenan de acuerdo con la cantidad de características que poseen.
2. Diferencias iguales en la característica representan diferencias iguales en los números asignados a las clasificaciones.
3. El punto cero representa la ausencia de características y la razón entre dos números es significativa.
4. La razón entre valores es significativa.

La siguiente grafica resume las principales características de los diversos niveles de medición.



# 2. Capitulo 2. Descripción de datos: tablas de frecuencias, distribuciones de frecuencias y su representación grafica.

## 2.1. Construcción de una tabla de frecuencias.

La estadística descriptiva se encarga de organizar datos con el fin de mostrar la distribución general de estos y el lugar en donde tienden concentrarse, ademas de señalar valores de datos pocos usuales o extremos. El primer procedimiento que se emplea para organizar y resumir un conjunto de datos es una **tabla de frecuencias.**

**Tabla de frecuencias:** Agrupación de datos cualitativos en clases mutuamente excluyentes que muestra el numero de observaciones en cada clase.

Recordar que, una variable cualitativa es de naturaleza no numérica; es decir, que la información es clasificable en distintas categorías. No hay un orden particular en estas categorías. Por otro lado, están las variables cuantitativas son de índole numérica.

### Frecuencias relativas de clase.

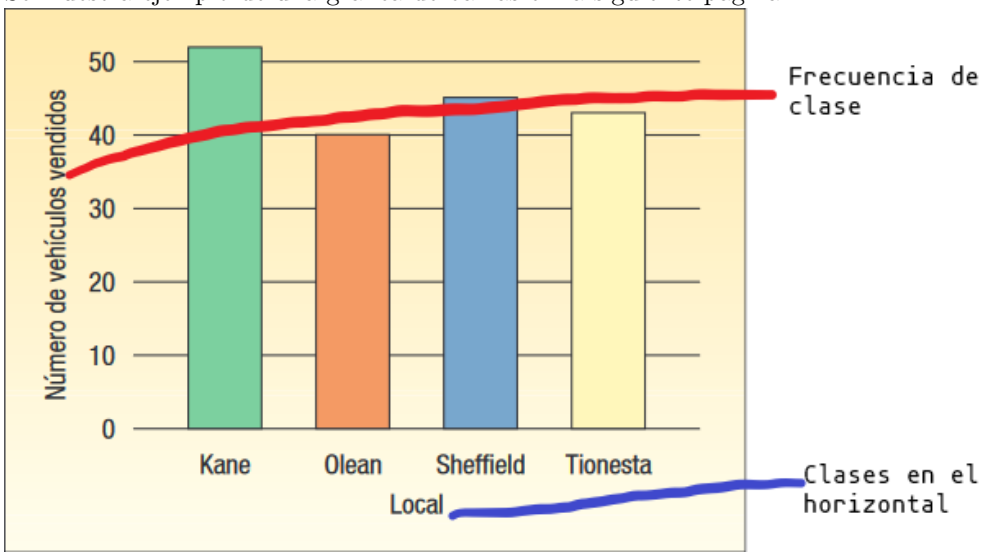
Es posible convertir las frecuencias de clase en frecuencias relativas de clase para mostrar la fracción del numero total de observaciones en cada una de ellas. Una frecuencia relativa capta la relación entre la totalidad de elementos de una clase y el numero total de observaciones. Para convertir una distribución de frecuencias en una distribución de frecuencias relativa, cada una de las frecuencias de clase se divide entre el total de observaciones.

### Representación grafica de datos cualitativos.

El instrumento mas común para representar una variable cualitativa en forma grafica es la **grafica de barras**. En la mayoría de los casos, el eje horizontal muestra la variable de interés y el eje vertical la frecuencia o fracción de cada uno de los posibles resultados. Una característica distinta de esta herramienta es que existe una distancia o espacio entre las barras. Una grafica de barras es una representación grafica de una tabla de frecuencias mediante una serie de rectángulos de anchura uniforme, cuya altura corresponde a la frecuencia de clase.

**Grafica de barras:** En ella, las clases se representan en el eje horizontal y la frecuencia de clase en el eje vertical. Las frecuencias de clase son proporcionales a las alturas de las barras.

Se muestra ejemplo de una grafica de barras en la siguiente pagina:

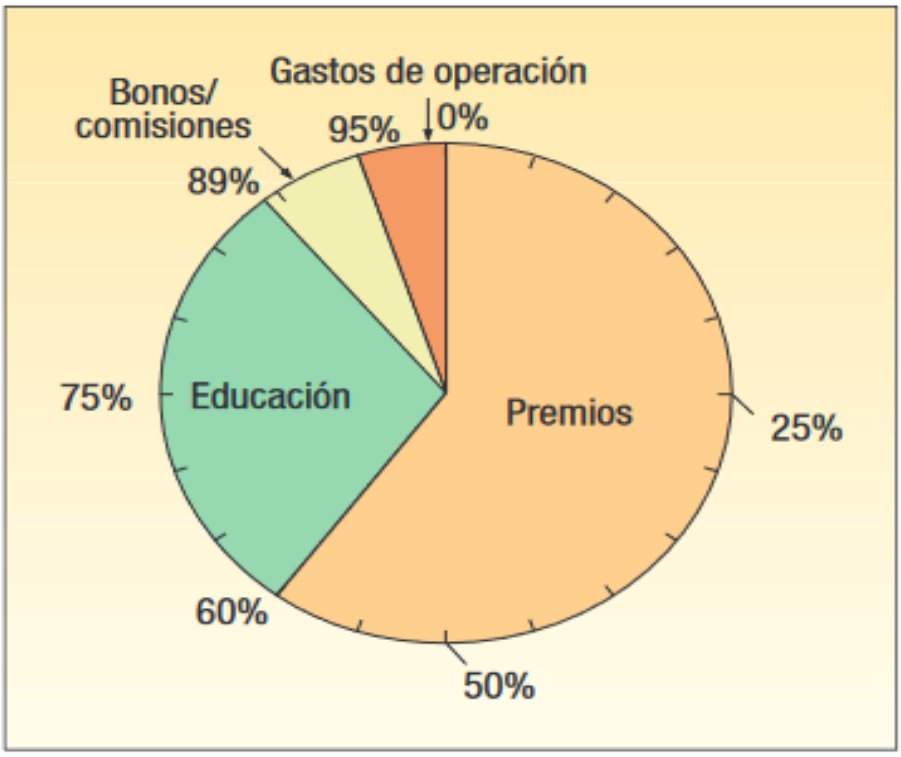



Otro tipo de grafica útil para describir información cualitativa es la **grafica de pastel**.

**Grafica de pastel:** Grafica que muestra la parte o porcentaje que representa cada clase del total de números de frecuencia.

El primer paso para elaborar una grafica de pastel consiste en registrar los porcentajes 0, 5, 10, 15, etc, de manera uniforme alrededor de la circunferencia de un circulo. El área rebanada representa alguna clase, cada rebanada de pastel representa la porción relativa de cada componente, es posible compararlas con facilidad.
Acá un ejemplo de grafica bizcochuelo.

| Uso del dinero de las ventas | Cantidad (millones de dólares) | Porcentaje de ventas |
|---|---|---|
| Premios | 1 460.0 | 60 |
| Educación | 702.3 | 29 |
| Bonos | 150.0 | 6 |
| Gastos | 124.3 | 5 |
| Total | 2 436.6 | 100 |




Las graficas de pastel y las de barras cumplen casi la misma función. ¿Cuales son los criterios para elegir una u otra? En la mayoría de los casos, las graficas de pastel son las mas informativas cuando se trata de comparar la diferencia relativa en el porcentaje de observaciones de cada uno de las variables de la escala nominal. Es preferible usar una grafica de barras cuando el objetivo es comparar el numero de observaciones en cada categoría.

## 2.2. Construcción de distribuciones de frecuencias: datos cuantitativos.

Distribución de frecuencias puede ser útil para describir la ganancias de ventas.

**Distribución de frecuencias:** Agrupación de datos en clases mutuamente excluyentes, que muestra el numero de observaciones que hay en cada clase.

¿Como crear una distribución de frecuencias? El primer paso consiste en acomodar los datos en una tabla que muestre las clases y el numero de observaciones que hay en cada clase. Recordá que el **objetivo** es construir tablas, diagramas y graficas que revelen rápidamente la concentración, los valores extremos y la distribución de los datos. La información desorganizada como **datos en bruto** o **datos no agrupados**. Los datos en bruto se interpretan con mayor facilidad si se organizan como una distribución de frecuencias.

**Paso 1: Defina el numero de clases.** El objetivo consiste en emplear suficientes agrupamientos o **clases**, de manera tal que se perciba la forma de la distribución. Acá se necesita criterio. Una gran cantidad de clases o muy pocas podrían no permitir ver la conformación fundamental del conjunto datos. Una receta útil para determinar la cantidad de clases ($k$) es la regla de 2 a la $k$. Esta guiá sugiere que se elija el menor numero ($k$) para el numero clases, de tal manera que $2^k$ sea mayor que el numero de observaciones ($n$). Ejemplo $n = 180$ observaciones, si supone $k = 7$, lo cual significa que utilizara siete clases, entonces $2^7 = 128$, algo menos que las 180 observaciones. De ahí que 7 no represente suficientes clases. Si $k = 8$, entonces $2^8 = 256$, que es mayor 180. Por lo tanto, el numero de clases se recomienda es de 8.

**Paso 2: Determine el intervalo o ancho de clase.** El **intervalo** o **ancho de clase** debería ser el mismo para todas las clases. Todas las clases juntas deben cubrir por lo menos la distancia del valor mas bajo al mas alto de los datos. Expresado esto en una formula seria:

$$i \geq \frac{H - L}{k}$$

Por lo general el tamaño de intervalo se redondea a una cifra conveniente, tal como un múltiplo de 10 a 100. En las distribuciones de frecuencia son preferibles los intervalos de clase iguales. Sin embargo, en ciertos casos se necesita que no lo sean para evitar una gran cantidad de clases vacías, o casi vaciás.

**Paso 3: Establezca los limites de cada clase.** Este paso es importante para que sea posible incluir cada observación en una sola categoría. Esto significa que debe evitar la superposición de limites de clase confusos. Por ejemplo, clases como \$1300 - \$1400 y \$1400 - \$1500 no deberían emplearse porque no resulta claro si el valor de \$1400 pertenece a la primera o a la segunda clases. Las clases como \$1300 - \$1400 y \$1500 - \$1600 se emplean con frecuencia, aunque también pueden resultar confusas si no se conviene en redondear todos los datos de \$1450 o por arriba de esta cantidad a la segunda clase y los datos por debajo de \$1400 a la primera clase. Al redondear el intervalo de clase hacia arriba con el fin de obtener un tamaño conveniente de clase, se cubre un rango mas amplio que el necesario. Una directriz consiste en convertir el limite inferir de la primera clase en un múltiplo del intervalo de clase. A veces esto no es posible, pero el limite inferior por lo menos debe redondearse.

**Paso 4: Anote las veces que encuentre las observaciones en el intervalo en las clases. PONER IMAGEN PASO 4**

**Paso 5: Cuente el numero de elementos de cada clase.** El numero de elementos que hay en cada clase recibe el nombre de **frecuencia de clase.** Por ejemplo en numero de \$200 a \$600 hay 8 observaciones, y en la clase de \$600 a \$1000 hay 11 observaciones. Por lo tanto la frecuencia de clase de la primera es

de 8, mientras que en la segunda es de 11. Hay un total de 180 observaciones o frecuencias en todo el conjunto de datos. así que la suma de todas las frecuencias debe ser igual a 180. **PONER TABLA2-7**

Con frecuencia aparecerán otros dos términos: **punto medio de clase** e **intervalo de clase.** El punto medio, que se encuentra entre los limites inferiores de dos clases consecutivas, se calcula sumando los limites interiores de clases consecutivas y dividiendo el resultado entre dos. En caso de la imagen del paso 5, el limite de clase interior de la primera clase es de $200 y el siguiente limite es de $600. El punto medio de clase $400, que se calcula mediante la operación

$$\frac{\$600 + \$200}{2}.$$

El punto medio de $400 representa mejor. Para determinar el intervalo de clase, se resta el limite inferior de la clase del limite inferior de la siguiente clase, es decir

$$\$600 - \$200 = 400.$$

Cambien se puede determinar el intervalo de clase calculando la diferencia entre puntos medios consecutivos. El punto medio de la primera clase des de $400 y el punto medio de la segunda clase es de $800. La diferencia es de $400.

## 2.3. Distribución de frecuencias relativas

Si se requiere resumir las ventas del ultimo mes utilizando ganancia por venta; seria útil describir la ganancia de venta por medio de una distribución de frecuencias.

**Distribución de frecuencias:** Agrupación de datos en clases mutuamente excluyentes, que muestra el numero de observaciones que hay en cada clase.

¿Como crear una distribución de frecuencias? El primer paso consiste en acomodar los datos en una tabla que muestre las clases y el numero de observaciones que hay en cada clase. El objetivo es construir tablas, diagramas y graficas que revelen rápidamente la concentración, los valores extremos y la distribución de los datos. La información desorganizada como **datos en bruto** o **datos no agrupados.** Los datos en bruto se interpretan con mayor facilidad si se organizan como una distribución de frecuencias.

**Paso 1: Defina el numero de clases:** El objetivo consiste en emplear suficientes agrupamientos o **clases**, de manera tal que se perciba la forma de la distribución. Se necesita criterio. Una gran cantidad de clases o muy pocas podrían no permitir ver la conformación fundamental del conjunto de datos. Una receta útil para determinar la cantidad de clases (k) es la regla de 2 a la k. Esta guía sugiere que se elija el menor numero (k) para el numero de clases, de tal manera que

$$2^k$$

sea mayor que el numero de observaciones (n). Si $n = 180$, si supone que $k = 7$, lo cual significa que utilizara siete clases, entonces

$$2^7 = 128$$

, algo menos de 180. De ahí que 7 no represente suficiente clases. Si $k = 8$, entonces

$$2^8 = 256$$

, que es mayor a 180. Por lo tanto, el numero de clases que se recomienda es de 8.

**Paso 2: Determine el intervalo o ancho de clase:** El **intervalo o ancho de clase** debería ser el mismo para todas las clases. Todas las clases juntas deben cubrir por lo menos la distancia del valor mas bajo al mas alto de los datos. Expresado esto en una formula seria:

$$i \geq \frac{H - L}{k}$$

en la que la $i$ es el intervalo de clase; H, el máximo valor observado; L, el mínimo valor observado, y k, el numero de clases. En la practica, por lo general este tamaño de intervalo se redondea a una cifra conveniente, tal como un múltiplo de 10 a 100. Las distribuciones de frecuencia son preferibles los intervalos de clase iguales. Ciertos casos se necesita que no lo sean para evitar una gran cantidad de clases vacías, o casi vacías. Facilita la compresión de la información.

**Paso 3: Establezca los limites de cada clase:** Este paso es importante para que esa posible incluir cada observación en una sola categoría. Esto significa que debe evitar la superposición de limites de clase confusos. Ponele tenes una clase 200-400 y otra de 400-600 ¿el 400 va incluido en la primera clase o en la segunda clase? ACLARA ESO O NO ENTIENDO NADADORA.

**Paso 4: Anote las observaciones en las clases:** Se da un ejemplo en la imagen siguiente **AGREGAR TABLA DEL PASO 4**

**Paso 5: Cuente el numero de elementos de cada clase:** El numero de elementos que hay en cada clase recibe el nombre **frecuencia de clase.** En la clase \$200 a \$600 hay 8 observaciones, y en la clase de \$600 a \$1000 hay 11 observaciones. Por lo tanto, la frecuencia de clase de la primera clase es de 8, mientras que en la segunda es de 11. Hay un total de 180 observaciones o frecuencias en todo el conjunto de datos. Así que la suma de todas las frecuencias debe ser igual a 180. **PEGAR TABLA2-7** Las ventajas de condesar los datos de forma mas entendible y organizada compensa por mucho este desventaja. Con frecuencia aparecerán otros dos términos: **punto medio de clase e intervalo de clase**. El punto medio, que se encuentra los limites inferiores de dos clases consecutivos, se calcula sumando los limites inferiores de clases consecutivas y dividiendo el resultado entre dos. Para determinar el intervalo de clase, se resta el limite inferior de la clase del limite inferior de la siguiente clase.

## 2.4. Distribución de frecuencias relativas.

Convertir frecuencias de clase en frecuencias relativas de clase, igual que con los datos cualitativos, con el fin de mostrar la fracción del total de observaciones que hay en cada clase. Una distribución de frecuencias relativas convierte la frecuencia en un porcentaje. Para convertir una distribución de frecuencia en una distribución de frecuencia relativa, cada una de las frecuencias de las clases se divide entre el numero total de observaciones. A continuación dejo la tabla de como serian una frecuencia relativa. **PONER TABLA 2-8**

## 2.5. Representación grafica de una distribución de frecuencias.

Es frecuente que se necesite una vista rápida de las tendencias de las ventas, los precios de las acciones o costos de hospitalización. A menudo, estas tendencias se describen por medio de tablas y graficas. Tres herramientas que serán de utilidad para representar gráficamente una distribución de frecuencias son el histograma, el polígono de frecuencias y el polígono de frecuencias acumuladas.

### Histograma

Un histograma de una distribución de frecuencias basadas en datos cuantitativos se asemeja mucho a la grafica de barras, que muestra la distribución de datos cualitativos. Las clases se señalan en el eje horizontal y las frecuencias de clase en el eje vertical. Las frecuencias de clase se representan por medio de las alturas de las barras. Existe una importante diferencia como consecuencia de la naturaleza de los datos. Por lo general, los datos cuantitativos se miden con escalas continuas, no discretas. Por lo tanto, el eje horizontal representa todos los valores posibles y las barras se colocan de forma adyacente para que muestren la naturaleza continua de los datos.

**Histograma:** Grafica en la que las clases se señalan en el eje horizontal y las frecuencias de clase en el eje vertical. Las frecuencias de clase se representan por medio de las altura de las barras, que se dibujan de manera adyacente.

Mostraremos un ejemplo a continuación.

### Poligono de frecuencias

Un polígono de frecuencias también muestra la forma que tiene una distribución y es similar a un histograma. Consiste en segmentos de recta que conectan los puntos que forman las intersecciones de los puntos medios de clase y frecuencias de clase. El punto medio de cada clase se indica en una escala en el eje X y las frecuencias de clase en el Y. Recordá que el punto medio de clase es el valor localizado en el centro de una clase y representa los valores típicos de ella (limite superior - limite inferior). La frecuencia de clase es el numero de observaciones que hay en una clase particular. Para construir un polígono de frecuencias, hay

que desplazarse horizontalmente sobre la grafica al punto medio, y en seguida de manera vertical, la frecuencia de clase, donde se coloca un punto. Los valores X y de Y de este punto reciben el nombre de *coordenadas*. El proceso continua con todas las clases. Posteriormente, los puntos se conectan de manera ordenada. Es decir, que el punto que representa la clase mas baja se une al que representa la segunda clase y así en lo sucesivo. **PONER GRAFICA 2-5** Tanto el histograma como el polígono de frecuencias permiten tener una vista rápida de las principales características de los datos (máximos, mínimos, puntos de concentración, etc.). Aunque las dos representaciones tienen un propósito similar, el histograma posee la ventaja de que describe cada clase como un rectángulo, en el que la barra de altura de este representa el numero de elementos que hay en cada clase. El polígono de frecuencias, en cambio, tiene una ventaja con respecto al histograma. También permite comparar directamente dos o mas distribuciones de frecuencias **PONER GRAFICA 2-6 DE EJEMPLO**.

#### Distribuciones de frecuencia acumulativas

. Una distribución de frecuencias acumulativas muestra el numero o porcentaje de observaciones por debajo de valores dados, con representación gráfica de un **polígono de frecuencias acumulativas.** Como su nombre lo indica, una distribución de frecuencias acumulativas y un polígono de frecuencias acumulativas implican *frecuencias acumulativas*. La frecuencia acumulativa básicamente vas sumando todas las frecuencias de las clases. **PONER TABLA 2-9** Para trazar una distribución de frecuencias acumulativas, **se ubica el limite superior de cada clase en una escala a lo largo del eje X**, y **las correspondientes frecuencias acumulativas, a lo largo del eje Y**. Para incluir información adicional, gradué el eje vertical a la izquierda en unidades y el eje vertical a la derecha en porcentajes.

# 3. Capitulo 3. Descripción de datos: medidas numéricas.

## 3.1. Introducción

En este capitulo se presentan dos formas numéricas de describir datos cuantitativos: las **medidas de ubicación** [2] y las **medidas de dispersión.** A las medidas de ubicación a menudo se les llama **promedios**. El propósito de una medida de ubicación consiste en señalar el centro de un conjunto de valores.

[2] La razón por la cual se llama *medida de ubicación* es porque estas medidas estadísticas representan la posición o ubicación de los datos dentro de un conjunto de datos. Es decir, indican dónde se encuentra la mayoría de los valores de un conjunto de datos y cómo están distribuidos alrededor de esta posición central. Por ejemplo, la media aritmética es una medida de ubicación porque representa el valor central alrededor del cual se agrupan los demás valores en un conjunto de datos. La mediana, por otro lado, es una medida de ubicación porque representa el valor que divide el conjunto de datos en dos partes iguales, es decir, la mitad de los valores están por encima de la mediana y la otra mitad están por debajo.

Vos ya estas conoces el concepto de promedio (media aritmética), medida de ubicación que muestra el valor central de los datos. Si solo tomas las medidas de ubicación en cuenta de un conjunto de datos o si compara varios conjuntos de datos utilizando valores centrales, llegara a una conclusión incorrecta. Ademas de las medidas de ubicación, debe tomar en consideración la **dispersión** (con frecuencia se le llama *frecuencia variación* o *propagación*) de los datos. Para describir la dispersión considere el rango, la desviación media, la varianza y la desviación estándar. En principio se explican las medidas de ubicación. No existe una única medida de dispersión; de hecho existen varias. Consideraremos cinco:

1. La media aritmética.
2. La media ponderada.
3. La mediana.
4. La moda.
5. La media geométrica.

La medida aritmética es la medida de ubicación que mas se utiliza y que se publica con mayor frecuencia, por lo cual se le considerará como parámetro para una población y como estadístico para las muestras.

## 3.2. La media poblacional

La media poblacional es la suma de todos los valores observados en la población dividida entre el numero de valores de la población. Para determinar la media poblacional, aplique la siguiente formula con la siguiente formula y formula con símbolos matemáticos:

$$\mu = \frac{\sum X}{N}$$

Cada símbolo representa:

- $\mu$ representa la media poblacional; se trata de la letra minúscula griega mu.
- $N$ es la numera de valores en la población.
- $X$ representa cualquier valor particular.
- $\sum$ es la letra mayúscula griega sigma e indica la operación de suma.
- $\sum X$ es la suma de X valores en la población.

Cualquier característica medible de una población recibe el nombre de **parámetro**. La media de una población es un parámetro.

**Parámetro:** Característica de una población.

## 3.3. Media de una muestra

Con frecuencia se selecciona una muestra (Se menciono en el capitulo 1) de la población para estimar una característica especifica de la población. La *media* es la suma de los valores de la muestra, divididos entre el numero total de valores de la muestra. La media de una muestra se determina de la siguiente manera:

$$\overline{X} = \frac{\sum X}{n}$$

Cada símbolo representa:

- $\overline{X}$ es la media de la muestra; se lee como "X barra."
- $n$ es el numero de valores de la muestra.
- $X$ representa cualquier valor particular.
- $\sum$ es la letra mayúscula griega sigma e indica la operación de suma.
- $\sum X$ es la suma de X valores en la población.

La media de una muestra o cualquier otra medición basada en una muestra de datos recibe el nombre de **estadístico**.

**Estadístico:** Característica de una muestra.

## 3.4. Propiedades de la media aritmética

La media aritmética es una medida de ubicación muy utilizada. Cuenta con algunas propiedades importantes:

1. **Todo conjunto de datos de intervalo -o de nivel de razón- posee una media.** Recordá que en el capitulo 1 los datos de nivel de razón incluyen datos como edades, ingresos y pesos, y que la distancia entre los números es constante.
2. **Todos los valores se encuentran incluidos en el calculo de la media.**
3. **La media es única.** Solo existe una media en un conjunto de datos. Mas adelante en el capitulo que un promedio que podría aparecer dos o mas veces en un conjunto de datos.
4. **La suma de las desviaciones de cada valor de la media es cero.** La media es un punto equilibrio de un conjunto de datos.

La media tiene un punto débil. Recuerde que el valor de cada elemento de una muestra o población, se utiliza cuando se calcula la media. Si uno de estos valores son extremadamente grandes o pequeños comparadas con la mayoría de los datos, la media podría no ser un promedio adecuado para representar los datos. Es decir, la media se ve afectada en exceso por valores inusualmente grandes o pequeños.

## 3.5. Media ponderada

La media ponderada, que constituye un caso especial de la media aritmética, se presenta cuando hay varias observaciones con el mismo valor. Es una medida estadística que se utiliza para calcular el valor medio de un conjunto de datos, donde cada valor tiene un peso o una importancia específica. En la media ponderada, cada valor se multiplica por su peso y luego se divide por la suma de todos los pesos.

$$\overline{X_w} = \frac{\sum(wX)}{\sum w}$$

Observe que el denominador de una medida ponderada siempre es la suma de las ponderaciones.

## 3.6. Mediana

Ya se dijo que si los datos contienen uno o dos valores muy grandes o muy pequeños, la media aritmética no resulta representativa. Es posible describir el centro de dichos datos a partir de una medida de ubicación denominada **mediana**.

**Mediana:** Punto medio de los valores una vez que se han ordenado de menor a mayor o de mayor a menor.

Ordenas los datos y agarras el(o los) valor (valores) del medio (literalmente). ¿Como se determina la mediana en el caso de un numero par de observaciones? se ordenan las observaciones. En seguida, con el fin de obtener un único valor por convención, calcule la media de las dos observaciones medias. Así, en el caso de un numero par de observaciones, la mediana quizá no sea uno de los valores dados. Ejemplo ponele que tenes dos observaciones medias 5 y 7, la media de esos dos es 6. Las principales propiedades de la mediana son las siguientes:

- **No influyen en ella valores extremadamente grandes o pequeños.** Por consiguiente, la mediana es una valiosa medida de ubicación cuando dichos valores se presentan.
- **Es calculable en el caso de datos de nivel ordinal o mas altos.** Recordá que en el capitulo 1 que los datos de nivel ordinal pueden ordenarse de menor a mayor

La mediana se determina para cualquier nivel de datos, excepto los nominales.

## 3.7. Moda

La moda es otra medida de ubicación (promedio).

**Moda:** Valor de la observación que aparece con mayor frecuencia.

La moda es de especial utilidad para resumir datos de nivel nominal. Es posible determinar la moda para todos los niveles de datos: nominal, ordinal, de intervalo y de razón. La moda también tiene la ventaja de que no influyen en ella valores extremadamente grandes o pequeños. No obstante, la moda tiene sus desventajas, por las cuales se le utiliza con menor frecuencias que a la media o mediana. En el caso de muchos conjuntos de datos no existe la moda porque ningún valor se presenta mas de una vez. Puede ocurrir que un agrupamiento de datos tenga dos modas se denomina bi-modal (tiene dos modas).

## 3.8. Posiciones relativas de la media, la mediana y la moda.

Se trata de una distribución simétrica que también tiene forma de campana. Esta distribución *posee* la misma forma a cualquier lado del centro. Si el polígono estuviera doblado a la mitad, las dos mitades serian idénticas. En cualquier distribución simétrica, la moda, la mediana, y la media siempre son iguales. Hay distribuciones simétricas que no tienen forma de campana (En una distribución en forma de campana la media, la mediana y la moda son iguales). Si una distribución no es simétrica, o **sesgada**, (Una distribución sesgada no es simétrica.) la relación entre las tres medidas cambia. En una distribución con sesgo positivo la media aritmética es la mayor de las tres medidas. ¿Por que? Porque en ella influyen, mas que sobre la mediana o la moda, unos cuantos valores extremadamente altos. Por lo general, la mediana es la siguiente medida mas grande en una distribución de frecuencias con sesgo positivo. La moda es la menor de las tres medidas. Si la distribución tiene un sesgo muy pronunciado, la media no seria una mediana adecuada. La mediana y la moda serian mas representativas.
Por el contrario, si una distribución tiene un **sesgo negativo**, la media es la menor medida de las tres. Por supuesto, la media es sensible a la influencia de una cantidad extremadamente pequeña de observaciones. La mediana es mayor que la media aritmética y la moda es la mas grande de las tres medidas. De nuevo, si la distribución tiene un sesgo muy pronunciado, la media no se utilizaría para representar a los datos.

## 3.9. ¿Por que estudiar la dispersión?

Una medida de ubicación, como la media o la mediana, solo describe el centro de los datos. Desde este punto de vista resulta valiosa, pero no dice nada sobre la dispersión de los datos. Una medida de dispersión pequeña indica que los datos se acumulan con proximidad alrededor de la media aritmética. Por consiguiente, la media se considera representativa de los datos. Por el contrario, una medida grande de dispersión indica que la media no es confiable. Otra razón para estudiar la dispersión en un conjunto de datos consiste en comparar la propagación en dos o mas distribuciones. Una medida de dispersión sirve para evaluar la confiabilidad de dos o mas medidas de ubicación.

## 3.10. Medidas de dispersión

Consideraremos diversas medidas de dispersión. El rango se sustenta en los valores máximo y mínimo del conjunto de datos, es decir, solo se consideran dos valores. La desviación media, la varianza y la desviación estándar se basan en desviaciones de la media aritmética.

### Rango

La medida mas simple de dispersión es el **rango**. Representa la diferencia entre los valores máximo y mínimo de un conjunto de datos. En forma de ecuación:

Rango = Valor máximo - valor mínimo

El rango se emplea mucho en aplicaciones de control de procesos estadísticos (CPE), debido a que resulta fácil de calcular y entender.

## 3.11. Rango intercuartílico o rango intercuartil

El rango intercuartílico o rango intercuartil (RIC o IQR), se obtiene como la diferencia entre los cuartiles superior e inferior: RIC $= Q_3 - Q_1$ mide el rango del 50 % central de la distribución. Gráficamente, está asociado a la "caja" del diagrama de caja.

### Desviación media

Un problema que presenta el rango escriba en que parte de dos valores, el mas alto y el mas bajo, es decir, no los toma en cuenta a todos. La desviación media si lo hace; mide la cantidad media respecto de la cual los valores de una población o muestran varían. Expresado en forma de definición:

**Desviación media**: Media aritmética de los valores absolutos de las desviaciones con respecto a la media aritmética.

En el caso de una muestra, la desviación media, designada DM, se calcula mediante la formula:

$$DM = \frac{\sum |X - \overline{X}|}{n}$$

en donde:

- $X$ es el valor de cada observación.
- $\overline{X}$ es la media aritmética de los valores.
- $n$ es el numero de observaciones en la muestra.

- || indica el valor absoluto.

¿Por que ignorar los signos de las desviaciones de la media? De no hacerlo, las desviaciones positivas y negativas se compensarían con exactitud unas a otras y la desviación media siempre seria cero. Dicha medida (cero) resultaría un estadístico sin utilidad.
La desviación media posee dos ventajas, primero, incluye todos los valores de los cálculos. Recuerde que el rango solo incluye los valores máximo y mínimo. Segundo, es fácil de definir: es la cantidad promedio que los valores se desvían de la media. Sin embargo, su inconveniente es el empleo de valores absolutos. Por lo general, es difícil trabajar con valores absolutos, así que la desviación media no se emplea con tanta frecuencia como otras medidas de dispersión, como la desviación estándar.

## Varianza y desviación estándar

La **varianza** y la **desviación estándar** también se fundamentan en las desviaciones de la media. Sin embargo, en lugar de trabajar con el valor absoluto de las desviaciones, la varianza y la desviación estándar lo hacen con el cuadrado de las desviaciones.

> **Varianza**: Media aritmética de las desviaciones de la media elevadas al cuadrado.

La varianza es no negativa y es cero solo si todas las observaciones son las mismas. El principal problema de la varianza es que se mide el términos del cuadrado de las unidades originales de medición. Por ejemplo:
Si las mediciones están dadas en metros, entonces la varianza queda expresada en metros cuadrados. Tomando la raíz cuadrada de la varianza, obtenemos la desviación estándar, que regresa la medida de variabilidad a la unidades de medición originales

> **Desviación estándar**: Raíz cuadrada de la varianza.

La **desviación estándar** de un conjunto de mediciones es igual a la raíz cuadrada de la varianza.
Desviación estándar poblacional: $\delta = \sqrt{\delta^2}$
Desviación estándar muestral: $s = \sqrt{s^2}$

## 3.12. Interpretación y usos de la desviación estándar

La desviación estándar normalmente se utiliza como medida para comparar la dispersión de dos o mas conjuntos de observaciones.

## 3.13. Coeficiente de variación

El coeficiente de variación (CV) nos informa acerca de la dispersión relativa de un conjunto de datos o mediciones.
Caso mediciones muéstrales: $CV = \frac{S}{|\overline{x}|}$
Caso mediciones poblacionales: $CV = \frac{\delta}{|\overline{\mu}|}$
Principales características del CV:

1. El coeficiente de variación no posee unidades, es decir es adimensional.
2. Para su interpretación se puede expresar como porcentaje, teniendo en cuenta que puede superar el 100 %.
3. Se utiliza para comparar conjuntos de datos pertenecientes a poblaciones distintas.
4. Cuanto más grande es el coeficiente de variación más dispersos están los datos respecto de la media, por lo que la media va perdiendo representatividad.

### Teorema de Chebyshev

Ya se ha insistido en el hecho de que una desviación estándar pequeña de un conjunto de valores indica que estos se localizan cerca de la media. Por el contrario, una desviación grande revela que las observaciones se encuentran muy dispersas con respecto a la media. El pibe Chebyshev estableció un teorema que nos permite determinar la mínima porción de valores que se encuentran a cierta cantidad de desviaciones estándares de la media. El teorema del Cheby establece lo siguiente:

> **Teorema de Chebyshev**: En cualquier conjunto de observaciones (muestra o población), la proporción de valores que se encuentran a $k$ desviaciones estándares de la media es de por lo menos $1 - 1/k^2$, siendo $k$ cualquier constante mayor que 1.

### La regla empírica

El teorema Cheby se relaciona con cualquier conjunto de valores; es decir, que la distribución de valores puede tener cierta forma. Sin embargo, en cualquier distribución simétrica con forma de campana, es posible ser mas precisos en la explicación de la dispersión en torno a la media. Estas relaciones que implican la desviación estándar y la media se encuentran descritas en la **regla empírica**, a veces denominada **regla normal**. La regla empírica solo se aplica a distribuciones simétricas con forma de campana.

**Regla empírica**: En cualquier distribución de frecuencias simétrica con forma de campana, aproximadamente 68 % de las observaciones se encontraran entre mas y menos una desviación estándar de la media; cerca de 95 % de las observaciones se encontraran entre mas y menos dos desviaciones estándares de la media y, de hecho (99.7 %), estarán mas y metros tres desviaciones de la media.

Se ha observado que si una distribución es simétrica y tiene forma de campana, todas las observaciones se encuentran entre la media y menos tres desviaciones estándares.

## 3.14. Media y desviación estándar de datos agrupados

En la mayoría de los casos las medidas de ubicación, como la media y las medidas de dispersión, como la desviación estándar, se determinan utilizando valores individuales. Sin embargo, algunas veces solo se cuenta con la distribución de frecuencias y se desea calcular la media o la desviación estándar. La siguiente explicación explica como calcular la media y la desviación estándar a partir de datos organizados en una distribución de frecuencias. Hay que insistir en que una media o una desviación estándar de datos agrupados es una estimación de los valores reales correspondientes.

### Media aritmética

Para aproximar la media aritmética de datos organizados en una distribución de frecuencia, comience suponiendo que las observaciones en cada clase se representan a través del punto medio de la clase. La media de una muestra de datos organizados en una distribución de frecuencias se calcula de la siguiente manera:

**Media aritmética de datos agrupados**:

$$\overline{X} = \frac{\sum fM}{n}$$

donde:

$\overline{X}$ designa la media muestra.

$M$ es el punto medio de cada clase.

$f$ es la frecuencia en cada clase.

$fM$ es la frecuencia en cada clase multiplicada por el punto medio de la clase.

$\sum fM$ es la suma de estos productos.

$n$ es el numero total de frecuencias.

### Desviación estándar

Para calcular la desviación estándar de datos agrupados en una distribución de frecuencias, necesita ajustar ligeramente la formula. Pondere cada una de las diferencias cuadradas por el numero de frecuencias en cada clase. La formula de la **Desviación estándar, datos agrupados** es:

$$s = \sqrt{\frac{\sum f(M - \overline{X})^2}{n - 1}}$$

donde:

$s$ es el símbolo de la desviación estándar de la muestra.

$M$ es el punto medio de la clase.

$f$ es la frecuencia de clase.

$n$ es el numero de observaciones en la muestra.

$\overline{X}$ designa la media muestra.

## 3.15. Ética e informe de resultados

Esta aprendiendo a organizar, resumir e interpretar datos mediante la estadística, también es importante que comprenda esta disciplina con el fin de que se convierta en un consumir inteligente de información. En este capitulo aprendió la forma de calcular estadísticas descriptivas de naturaleza numérica. En particular, la manera de calcular e interpretar medidas de ubicación de un conjunto de datos: la media, la mediana y la moda. También ha estudiado las ventajas y desventajas de cada estadístico. Conocer las ventajas y desventajas de la media, la mediana y la moda es importante al dar un informe estadístico y cuando se emplea información estadística para tomar decisiones. También aprendió a calcular medidas de dispersión: el rango, la desviación media y la desviación estándar. Cada uno de estos estadísticos tiene ventajas y desventajas. Recuerde que el rango proporciona información sobre la dispersión total de una distribución. Sin embargo, no aporta información sobre la forma en que se acumulan los datos o se concentran entorno al centro de la distribución. Conforme aprenda más estadística, necesitara recordar que cuando emplea esta disciplina debe mantener un punto de vista independiente y basado en principios. Cualquier informe estadístico requiere la comunicación honesta y objetiva de los resultados.

# 4. Descripción de datos

En este capítulo continua el estudio de la estadística descriptiva. Estos diagramas y la estadística proporcionan una idea adicional del lugar en el que los

valores se concentran, así como de la forma general de los datos. En seguida se consideran datos bivariados de cada una de las observaciones individuales o seleccionadas.

> **Parámetros**: medidas descriptivas numéricas calculadas a partir de **mediciones poblacionales.** Por ejemplo, la media aritmética o promedio de una población, es decir, la media poblacional.
>
> **Estadísticos (o estadísticas)**: medidas descriptivas numéricas calculadas a partir de **mediciones muéstrales.** Por ejemplo, la media aritmética o promedio de una muestra, es decir, la media muestral

Los parámetros se designan con **letras griegas** y los estadísticos con **letras latinas**, a veces con algún añadido como el caso de la media (con una barra arriba de la letra).

## 4.1. Sesgo

Hay cuatro formas: simétrica, con sesgo positivo, con sesgo negativo y bimodal. En un conjunto **simétrico** de observaciones la media y mediana son iguales, y los valores de datos se dispersan uniformemente en torno a estos valores. Un conjunto de valores se encuentra **sesgado a la derecha** o **positivamente sesgado** o **sesgado a la izquierda** si existe un solo pico y los valores a se extienden mucho mas allá a la derecha del pico que a la izquierda de este. En este caso **la media es mas grande que la mediana**. En una distribución **negativamente sesgada** existe un solo pico, pero las observaciones se extienden mas a la izquierda, en dirección negativa. En una distribución negativamente sesgada, **la media es menor que la mediana**. Una **distribución bimodal** tendrá dos o mas picos. Con frecuencia este es el caso cuando los valores provienen de dos o mas poblaciones

## 4.2. Percentiles, deciles y cuartiles

La desviación estándar es la medida de dispersión que mas se utiliza. No obstante, existen otras formas de describir la variación o dispersión de un conjunto de datos. Un método consiste en determinar la ubicación de los valores que dividen un conjunto de observaciones en partes iguales. Estas medidas incluyen los **cuartiles, deciles y percentiles**.
Los **cuartiles** dividen a un conjunto de observaciones en cuatro partes iguales, ordenadas de menor a mayor. Para explicarlo mejor, piense en un conjunto de valores ordenados de menor a mayor. Anteriormente se denomino a la *mediana* al valor intermedio de un conjunto de datos ordenados de menor a mayor. Es decir el 50 % de las observaciones son mayores que la mediana y 50 % son menores que la mediana. La mediana constituye una medida de ubicación, ya que señala el centro de los datos. De igual manera, los **cuartiles** dividen a un conjunto de observaciones en cuatro partes iguales. El primer cuartil, que se representa con $Q_1$, es el valor debajo del cual se presenta el 25 % de las observaciones, y el

tercer cuartil, que simboliza $Q_3$, es el valor debajo del cual se presenta 75 % de las observaciones. Lógicamente, $Q_2$ es la mediana. $Q_1$ puede considerarse como la mediana de la mitad inferior y $Q_3$ como la mediana de la parte superior de los datos.
Asimismo, los **deciles** dividen un conjunto de observaciones en 10 partes iguales y los **percentiles** en 100 partes iguales.
Un ejemplo de deciles es, si su promedio general en la universidad se encuentra en el octavo decil, usted podría concluir que 80 % de los estudiantes tuvieron un promedio general inferior al suyo y 20 %, un promedio superior.
Un ejemplo de percentiles es, un promedio general ubicado en el trigésimo tercer percentil significa que el 33 % de los estudiantes tienen un promedio general mas bajo y que el 67 % un promedio mas alto.

## 4.3. Diagramas de caja

Un **diagrama de caja** o **box-plot** es una representación grafica, basada en cuartiles, que ayuda a presentar un conjunto de datos. Para construir un diagrama de caja, solo necesita cinco estadísticos:

1. El valor mínimo.
2. Primer cuartil.
3. La mediana (segundo cuartil).
4. Tercer cuartil.
5. El valor máximo.

### Construcción de un diagrama de caja

Cualquier medición a mayor distancia del límite superior o mayor distancia del límite inferior es un resultado o valor atípico; el resto de las mediciones, dentro de los límites, no son inusuales. Los resultados atípicos se suelen marcar con un asterisco o punto. El diagrama de caja marca el rango del conjunto de datos usando "bigotes" para conectar las mediciones más pequeñas y más grandes (excluyendo resultados atípicos) a la caja.

**Límite inferior**: $Q_1 - 1,5.(Q_3 - Q_1)$

**Límite superior**: $Q_3 + 1,5.(Q_3 - Q_1)$

**Importante:** Los límites son imaginarios, no se suelen trazar A la diferencia $Q_3 - Q_1$ se le llama **rango intercuartílico** o **rango intercuartil**. Un diagrama de caja puede representarse de manera horizontal o vertical. Suele ser más sencillo de interpretar de manera horizontal.

# 5. Probabilidad

A la estadística descriptiva le concierne el resumen de datos recogidos de eventos pasados. Ahora se presenta la segunda faceta de la estadística, a saber, el *calculo de la probabilidad de que algo ocurra en el futuro*. Esta faceta de la estadística recibe el nombre de **inferencia estadística** o **estadística inferencial**. La inferencia estadística se relaciona con las conclusiones relacionadas con una población sobre la base de una muestra que se toma de ella. Dada la incertidumbre existente en la toma de decisiones, es importante que se evalúen científicamente todos los riesgos implicados. La teoría de la probabilidad, a menudo conocida como la ciencia de la incertidumbre, resulta útil para hacer esta evaluación. Su aplicación permite a quien toma decisiones y posee información limitada analizar los riesgos y reducir al mínimo el riesgo que existe, por ejemplo, al lanzar al mercado un nuevo producto o aceptar un envió que contenga partes defectuosas. Puesto que los conceptos de la probabilidad son importantes en el campo de la inferencia estadística, en este capitulo se introduce el lenguaje básico de la probabilidad, que incluye términos como experimento, evento, probabilidad subjetiva y reglas de la adición y de la multiplicación.

## 5.1. ¿Que es la probabilidad?

En general es un numero que describe la posibilidad de que algo suceda.

> **Probabilidad**: Valor entre cero y uno, inclusive, que describe la posibilidad relativa (oportunidad o casualidad) de ocurra un evento.

Es común que una probabilidad sea expresada en forma décima, como 0.70, 0.27 o 0.50. No obstante, también se da forma de fracción, como 7/10, 27/100 o 1/2. Se puede suponer cualquier numero de 0 a 1, inclusive. La probabilidad de 1 representa algo que seguramente sucederá, y la probabilidad de 0 representa algo que no sucederá. Cuanto mas próxima se encuentre una probabilidad a 0, mas improbable es que el evento suceda. Cuanto mas próxima se encuentre la probabilidad a 1, mas seguro que suceda. En el estudio de la probabilidad se utilizan tres palabras clave: **experimento, resultado** y **evento**.

> **Experimento**: Proceso que induce a que ocurra una y solo una de varias posibles observaciones. Es el proceso mediante el cual se obtiene un resultado (observación o medición)

Respecto de la probabilidad, un experimento tiene dos o mas posibles resultados y no sabe cual ocurrirá. Un experimento es aleatorio cuando no se puede predecir el resultado que se va a obtener. Por ejemplo, lanzar una moneda, tirar un dado, jugar a la lotería, realizar una encuesta, etc.

> **Resultado**: Resultado particular de un experimento.

Un ejemplo: lanzar una moneda al aire constituye un experimento. Usted puede observar el lanzamiento de una moneda, pero no esta seguro si caerá *cara* o

*cruz*. Si se lanza una moneda, un resultado particular es *cara*. El otro posible resultado es *cruz*. Cuando se observan uno o mas resultados en los experimentos constituyen un evento.

> **Evento**: Conjunto de uno o mas resultados de un experimento. Se lo suele denotar con una letra mayúscula.

Un ejemplo: En el caso del experimento del lanzamiento de un dado, hay seis posibles resultados, pero existen varios posibles eventos. Cuando se cuenta el numero de miembros de la junta directiva de las compañías de Fortune 500 que tienen mas de 60 años de antigüedad el numero posibles de resultados varia de cero al total de miembros. Hay un numero aun mayor de eventos posibles en este experimento. **PONER TABLA PAGINA 147, CAP 5.2 ¿que es la probabilidad?**

## 5.2. Enfoques para asignar probabilidades.

**Probabilidad clásica.**

La **probabilidad clásica** parte del supuesto de que los resultados de un experimento son igualmente posibles. De acuerdo con el punto de vista clásico, la probabilidad de un evento que se esta llevando a cabo se calcula dividiendo el numero de resultados favorables entre el numero de posibles resultados.

> **Probabilidad clásica**:
>
> $$\text{Probabilidad de un evento} = \frac{\text{Numero de resultados favorables}}{\text{Numero total de posibles resultados}}$$

El concepto de conjuntos mutuamente excluyentes, recordá que cuando creamos clases de tal manera que un evento particular se incluyera en una sola de las clases y que no hubiera superposición entre ellas. Por lo tanto, solo uno de varios eventos puede presentarse en cierto momento.

> **Mutuamente excluyente**: El hecho de que un evento se presente significa que ninguno de los demás eventos puede ocurrir al mismo tiempo.

Unos ejemplos: La variable *genero* da origen a resultados mutuamente excluyentes: hombre y mujer. Un empleado seleccionado al azar es hombre o mujer, pero no puede tener ambos géneros. Una pieza fabricada es aceptable o no lo es. La pieza no puede ser aceptable e inaceptable al mismo tiempo. En una muestra de piezas fabricadas, el evento de seleccionar una pieza no aceptable y el evento de seleccionar una pieza aceptable son mutuamente excluyentes. Si un experimento incluye un conjunto de eventos con todo tipo de resultados posibles, como los eventos ün numero parz ün numero impar.$^{\text{en}}$ el experimento del lanzamiento del dado, entonces el conjunto de eventos es **colectivamente exhaustivo**. En el experimento del lanzamiento del dado, cada resultado sera par o impar. Por consiguiente, el conjunto es colectivamente exhaustivo.

> **Colectivamente exhaustivo**: Por lo menos uno de los eventos debe ocurrir cuando se lleva a cabo un experimento.

Si el conjunto de eventos es colectivamente exhaustivo y los eventos son mutuamente excluyentes, la suma de las probabilidades es 1. Resulta innecesario llevar a cabo un experimento para determinar la probabilidad de un evento mediante el enfoque clásico, ya que el numero total de resultados se sabe antes de realizar el experimento. Por lógica, es posible determinar la probabilidad de sacar una cruz al lanzar una moneda o tres caras al lanzar tres monedas.

**Probabilidad empírica.**

La **probabilidad empírica** o **frecuencia relativa**, el segundo tipo de probabilidad, se basa en el numero de veces que ocurre el evento como proporción del numero de intentos conocidos.

> **Probabilidad empírica**: La probabilidad de que un evento ocurra representa una fracción de los eventos similares que sucedieron en el pasado.

En términos de una formula:

$$\text{Probabilidad empírica} = \frac{\text{Numero de veces que el evento ocurre}}{\text{Numero total de observaciones}}$$

El enfoque empírico de la probabilidad se basa en la llamada *ley de los grandes números*. La clave para determinar probabilidades de forma empírica consiste en que una mayor cantidad de observaciones proporcionaran un calculo mas preciso de la probabilidad.

> **Ley de los grandes números**: En una gran cantidad de intentos, la probabilidad empírica de un evento se aproximara a su probabilidad real.

Para explicar la ley de los grandes números, supongamos que lanzamos una moneda común. El resultado de cada lanzamiento es cara o cruz. Si se lanza la moneda una sola vez, la probabilidad empírica de las caras es cero o uno. Si lanzamos la moneda una gran cantidad de veces, la probabilidad del resultado de las caras se aproximara a 0.5. Observa que conforme incrementamos el numero de intentos, la probabilidad empírica de que salga una cara se aproxima a 0.5, que es su valor de acuerdo con el enfoque clásico de la probabilidad. ¿Que demostramos? Que a partir de la definición clásica de probabilidad, la posibilidad de obtener una cara en un solo lanzamiento de una moneda común es de 0.5. Según el enfoque empírico de la frecuencia relativa de la probabilidad, la probabilidad del evento se aproxima al mismo valor determinado de acuerdo con la definición clásica de probabilidad. Este razonamiento permite emplear el enfoque empírico y de la frecuencia relativa para determinar una probabilidad.

## 5.3. Reglas para calcular probabilidades

Ahora, que definimos probabilidad y descrito sus diferentes enfoques, cabe atender al calculo de la probabilidad de dos o mas eventos aplicando las reglas de la adición y multiplicación.

### Regla general de la adición

Cuando dos eventos ocurren al mismo tiempo, la probabilidad se denomina **probabilidad conjunta**.

> **Probabilidad conjunta**: Probabilidad que mide la posibilidad de que dos o mas eventos sucedan simultáneamente.

Esta regla para dos eventos designados A y B se escribe:

Regla general de la adición P(A o B) = P(A) + P(B) - P(A y B)

En el caso de la expresión P(A o B), la conjunción o sugiere que puede ocurrir A o puede ocurrir B. Esto también incluye posibilidad de que A y B ocurran. Tal uso o a veces se denomina **inclusivo**. También es posible escribir P(A o B o Ambos) para hacer hincapié en el hecho de que la unión de dos eventos incluye la intersección de A y B. Si comparamos las reglas general y especial de la adición, la diferencia que importa consiste en determinar si los eventos son mutuamente excluyentes. Si lo son, entonces la probabilidad conjunta P(A y B) es 0 y podríamos aplicar la regla especial de la adición. De lo contrario, debemos tomar en cuenta la probabilidad conjunta y aplicar la regla general de la adición.
Por ejemplo, si tenemos los eventos:
A: obtener un 1 o un 4. En símbolos: A = {1,4}
B: obtener un numero par. En símbolos: B = {2,4,6} Entonces:

$$P(A \cup B) = \frac{2}{6} + \frac{3}{6} - \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \text{ por lo que: } P(A \cup B) = \frac{2}{3}$$

### Regla especial de la adición

Para aplicar la regla especial de la adición, los eventos deben ser *mutuamente excluyentes*. Recordá que significa que cuando un evento ocurre, ninguno de los otros eventos puede ocurrir al mismo tiempo. Si dos eventos A y B son mutuamente excluyentes, la regla especial de la adición establece que la probabilidad de que ocurra uno u otro es igual a la suma de sus probabilidades. Esta regla se expresa mediante la siguiente formula:

**Regla especial de la adición**: P(A o B) = P(A) + P(B)

En el caso de los tres eventos mutuamente excluyentes designados A, B y C, la regla se expresa de la siguiente manera:

P(A o B o C) = P(A) + P(B) + P(C)

**Regla del complemento**

**Regla general de la multiplicación**

**Regla especial de la multiplicación**

## 5.4. Principios de conteo

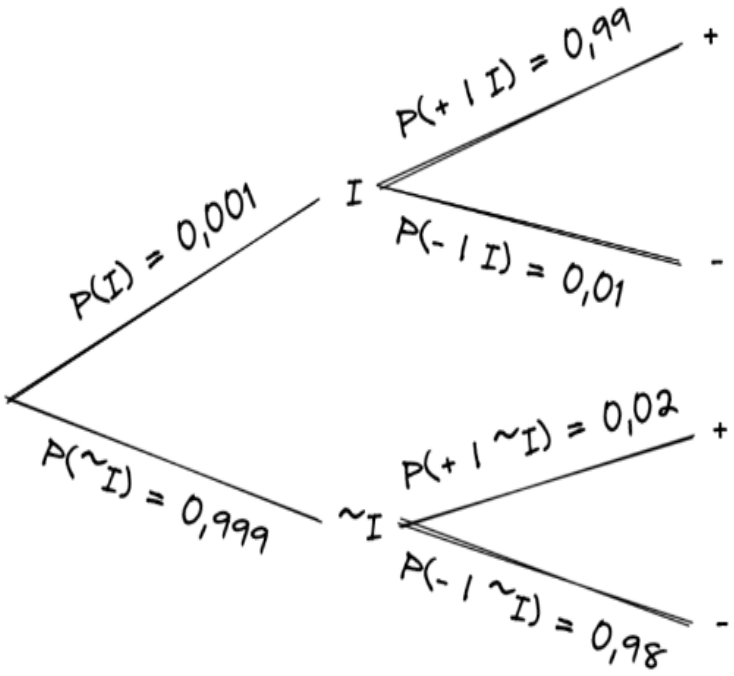
## 5.5. Teorema de Bayes

El teorema de Bayes es un caso de aplicación de la probabilidad condicional, el cual se puede aplicar cuando se cuenta con información a priori de los datos que se están analizando. Pero para comenzar a introducirnos a este teorema primero tenemos que tener en cuenta una serie de cosas.

1. Tener conocimiento de que es un diagrama de árbol, como leerlos y como este nos va a ayudar en el planteamiento del problema.
2. El concepto de probabilidad total de un evento.
3. La regla general de la multiplicación, que se utiliza para obtener la fórmula del teorema de Bayes.

### Diagrama de Árbol

Los diagramas de árbol, son una forma de representar las distintas condiciones o alternativas a las que están sometidas las variables de un problema, para su mejor comprensión supongamos que nos encontramos con el siguiente problema: El 0,1 % de la población de un país está infectada con una bacteria, y la prueba para detectar la enfermedad tiene una efectividad del 99 % en las personas enfermas y para las personas que no están infectadas este porcentaje es de 2 %. Con fin de representar estos datos de forma más gráfica y ordenada, podemos poner toda esta información en un diagrama de árbol, el cual quedaría conformado de la siguiente manera.




Como se puede observar cada rama del árbol contiene los respectivos valores representativos para cada situación. Un dato importante a tener en cuenta a la hora de realizar este tipo de diagrama es que la suma de las probabilidades de las ramas pertenecientes al mismo padre deben dar siempre 1, por ejemplo, si la probabilidad de estar infectado es de 0,001 entonces la de no estar infectado debe ser si o si 0,999; ya que la sumas de estos dos siempre tiene que dar 1.

**Teorema de la Probabilidad Total**

Siguiendo con el ejemplo presentado anteriormente, supongamos que queremos saber cual es la probabilidad de que una persona de positiva en la prueba, independientemente de si está infectado con la bacteria o no. En ese caso se tendrán que sumar todos los caminos posibles del diagrama de árbol que nos lleve hacia ese evento específico. Aunque ya existe una fórmula que nos facilita encontrar la probabilidad total.

$$\sum_{j=1}^{k} P(S_j)P(A|S_j)$$

Nótese que en sí esta ecuación en primera instancia es la sumatoria de las intersecciones entre los demás eventos del experimento, con el evento que nosotros estamos analizando, en nuestro caso cuál es la probabilidad de que de positivo

una prueba, independientemente de si está infectado o no. Dicha sumatoria nos quedaría de la siguiente manera.

$$\sum_{j=1}^{k} S_j \cap A$$

Esta fórmula se transforma, en la del teorema de la probabilidad total, si aplicamos la regla general de la multiplicación, en la cual una intersección entre eventos se puede representar como la multiplicación de la siguiente manera:

$$P(S \cap A) = P(S)P(A|S)$$
$$P(S \cap A) = P(A)P(S|A)$$

**Fórmula del Teorema de Bayes**

Supongamos ahora que queremos saber la probabilidad de que la prueba de positiva siendo que estas infectado con la bacteria. En ese caso ese valor sería obtenido a través de la siguiente fórmula:

$$P(S_i|A) = \frac{P(S_i)P(A|S_i)}{\sum_{j=1}^{k} P(S_j)P(A|S_j)}$$

Como podemos ver, esta fórmula es la misma que para la probabilidad condicionada, con la diferencia que se realiza una serie de cambios leves, resultado de aplicar las propiedades antes mencionadas. Este teorema resulta muy útil para muchos campos de la ciencia. Pero tiene una serie de puntos débiles que lo convierten en una no muy buena opción para algunas situaciones en particular.

1. Si no se tiene un conocimiento del fenómeno, no es posible asignar las probabilidades iniciales (a priori) y por lo tanto no es posible aplicar el teorema de Bayes. Este problema se suele evitar, suponiendo que las probabilidades a priori tienen una distribución uniforme, lo cual es considerado por algunos autores como inaceptable, porque en muchas situaciones no conocemos la distribución inicial, pero sabemos claramente que no es uniforme. Otra posibilidad sería asignar subjetivamente las probabilidades iniciales, es decir asignar probabilidad sin contar con evidencias que la respaldan, pero esto llevaría a que dos investigadores con los mismos datos obtuviesen unas probabilidades finales diferentes.

2. El método de Bayes no permite calcular ni revisar "*probabilidades a priori objetivas*"de la hipótesis sino las "*probabilidades a priori subjetivas*"de dicha hipótesis, la cual es establecida por cada investigador. Es decir su grado personal de creencia en la veracidad de la hipótesis.