UNIDAD 3: LOS VALORES QUE CARACTERIZAN AL CONJUNTO DE DATOS

1. ¿Por qué son Necesarios?

En el Capítulo anterior hemos analizado herramientas estadísticas elementales que permiten resumir grandes masas (conjuntos) de datos primarios (categóricos o numéricos), convirtiéndolos en expresiones comprensibles y operables como lo son las tablas y los gráficos de las distribuciones de frecuencias. Además, hemos introducido algunas medidas simples que ayudan a la interpretación de tales resúmenes: frecuencias relativas y acumuladas.

La correcta utilización de esas herramientas descriptivas nos permitirá elaborar ciertas conclusiones sobre los "individuos" observados. Por ejemplo, analizando las tablas y gráficos del capítulo anterior, en las que se resumen diferentes grupos de datos relativos a los estudiantes del Curso de Estadística, podríamos afirmar entre otras cosas que[1]:

- ✓ *el 13% de los alumnos dedica 3 horas diarias a mirar TV*,
- ✓ *109 alumnos del curso son mujeres*,
- ✓ *90 estudiantes tienen 24 años o menos*.

A menudo, el análisis y descripción que deseamos realizar requiere de medidas capaces de ***resumir*** aún más al conjunto de datos, expresándolo en ***un solo "valor"*** (número o categoría de la variable en estudio) que lo ***represente***. Expresiones de síntesis como las siguientes facilitarán la comprensión global del fenómeno que expresan los datos que se analizan y, además, harían más sencilla la comparación entre distintas series de datos:

- ✓ *"Los grupos turísticos registran una estadía **promedio de 3 noches** en Puerto Iguazú"*.
- ✓ *"Es llamativo que el 50 por ciento de los usuarios de la red tiene **más de 50 años**"*.
- ✓ *"El **fresno** es el árbol que **más abunda** en la ciudad de Buenos Aires, con más del 40% del total de ejemplares*".

En los tres ejemplos, cada uno de los conjuntos de datos analizados (pernoctes en Puerto Iguazú, edad de los usuarios de Internet y variedad de los árboles de la CBA), queda ***resumido y expresado*** por un único valor de la variable en estudio: *"3 noches"*, *"50 años"* y *"fresno"*. Estas son las medidas estadísticas denominadas *"de tendencia central"*.

IMPORTANTE

Es oportuno reiterar que las medidas presentadas en el Capítulo anterior (frecuencias absolutas, relativas, etc.) y las que veremos en esta unidad, se **emplean de igual modo y con idénticos fines** de **resumen y descripción**, ya sea cuando se trata de **datos muestrales** como de **datos poblacionales** ("censales"). Es decir que, tanto los **conceptos** como la **forma de calcularlas** y la **interpretació**n de los resultados, son los mismos en ambas situaciones de trabajo.

En Capítulos posteriores distinguiremos el significado que adquieren estas medidas (estadístico muestral/estimador o parámetro) según provengan de datos muestrales o poblacionales.

[1] Sugerimos que el lector identifique las medidas estadísticas utilizadas en cada una de estas afirmaciones y que, aplicándolas a los datos de los ejemplos citados, verifique que todas ellas sean correctas.

2. ¿Cuáles Son?

Las medidas de tendencia central de un conjunto de datos son valores que ***tienden a ubicarse en el centro de la distribución*** (de ahí su nombre), cuando esta reúne ciertas condiciones: es unimodal [2] y la mayor concentración de los datos (mayores frecuencias) ocurre alrededor de los valores centrales de la variable observada.

Son varias las medidas de resumen llamadas de tendencia central: las que se construyen mediante ***alguna forma*** (aritmética, geométrica, cuadrática o armónica) de ***promediar todos los datos*** del conjunto y las que se ***basan en un solo dato*** de la serie (mediana y modo). En este curso analizaremos solo las tres de uso más común:

- el promedio aritmético o *"media aritmética",*
- la *moda o modo*, y
- la *mediana.*

IMPORTANTE

A lo largo del texto iremos introduciendo la notación matemática ("fórmulas") de las herramientas estadísticas que analizaremos y, en ciertos casos, de algunas demostraciones relacionadas con ellas.

Como regla general, estas expresiones estarán a continuación del concepto estadístico que representan. Por ello, ***recomendamos firmemente*** *centrar la atención y asegurarse de* ***comprender primero el concepto,*** *luego su formalización matemática, y por último el procedimiento de cálculo.*

3. Media Aritmética

Concepto

La media aritmética $\bar{x}$ de un conjunto de datos de una **variable numérica** "X", es el resultado de **sumar todos** los valores del conjunto y **dividir esa suma** por el total **n** de observaciones que componen el conjunto [3].

Simbología: La notación usual para representar a la media aritmética es: $\bar{x}, \bar{y}, \bar{z}$, etc., dependiendo de la letra (X, Y ó Z) adoptada para simbolizar a la variable en estudio. La distinción entre letras mayúsculas ($\overline{X}$) y minúsculas ($\bar{x}$) generalmente se reserva para diferenciar una media poblacional (mayúscula) de una muestral (minúscula). En este curso ***utilizaremos única e indistintamente*** la notación $\bar{x}$, debiendo el lector tener presente la advertencia anterior.

De igual modo, las letras *n* y *N* son usualmente reconocidas para distinguir en forma simbólica al total de observaciones de una muestra (*n*) y al total de datos de una población (*N*). Utilizaremos el símbolo *n* indistintamente.

Así entonces, si tomáramos los $n = 136$ datos [4] de la variable "Y" (columna) "*edad*", registrados en la matriz *"Estudio de los Alumnos de Estadística I"* del Capítulo anterior, el ***promedio*** o ***media aritmética*** o simplemente ***"media"*** de ese conjunto de observaciones, será:

$$\bar{y} = \frac{19 + 27 + 26 + 28 + \ldots\ldots\ldots + 30}{136} = \frac{3180}{136} \cong 23,4 \text{ años}$$

`valor promedio o "media aritmética" del conjunto`

`total de datos`

[2] El concepto de distribución "unimodal" quedará debidamente aclarado en puntos posteriores de esta unidad.

[3] Nótese que por tratarse de una medida "calculada" con los datos, ***solo es aplicable*** a datos de ***variables numéricas***.

[4] No declaran su edad 3 estudiantes.

Vemos en el ejemplo cómo la media aritmética resume en un solo número toda la información del conjunto de individuos observados: "*se trata de un grupo de 136 estudiantes cuya edad promedio es de, aproximadamente, 23 años"*.

Actividad Nº 1

Antes de continuar con la lectura, deberá realizar aquí la Actividad Nº 1 de la Guía de Actividades correspondiente a esta unidad.

En Fórmula

Sea $\{x_1, x_2, x_3, x_4, x_5, \ldots .. x_i, \ldots\ldots x_n\}$; un conjunto de *n* observaciones de la ***variable numérica*** "X". Según la definición anterior, el valor $\bar{x}$, promedio o media aritmética del conjunto, será:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_i + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Notaciones Equivalentes

Otras formas matemáticas equivalentes para expresar al promedio, son las siguientes:

$$\bar{x} = \sum \frac{xi}{n} \qquad\qquad \bar{x} = \frac{1}{n} \sum xi$$

3.1. Principales Propiedades de $\bar{x}$

La media aritmética reúne ciertas propiedades que es importante conocer para utilizarla correctamente como resumen de un conjunto de datos, o bien para resolver algunos problemas que pueden surgir en su aplicación práctica.

- ***Primera Propiedad***

  Si dos de los términos de la expresión $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ son conocidos, se puede determinar el tercero de ellos mediante un simple pasaje de términos. Cuando se conocen $\bar{x}$ y *n,* la suma $\sum_{i=1}^{n} x_i$ se podrá determinar haciendo el producto de $\bar{x}$ por *n*. En símbolos:

  $$\sum_{i=1}^{n} x_i = \bar{x} \,.\, n$$

  Esta propiedad matemática nos permitiría saber, por ejemplo, que las $n = 32$ cárceles federales [5] de todo el país alojan un total de 60.416 internos, ya que cada una de ellas tiene una media de 1.888 presos. Esto es así porque:

  $$\sum_{1}^{32} x_i = 32 \,.\, 1888 = 60416$$

- ***Segunda Propiedad***

  El promedio es una ***medida calculada*** a partir de todos y cada uno de los datos de una serie, en consecuencia resume apropiadamente la información del conjunto. Sin embargo, por esta propiedad, en ciertas situaciones de trabajo puede perder eficacia como medida *"representativa"* del conjunto de datos.

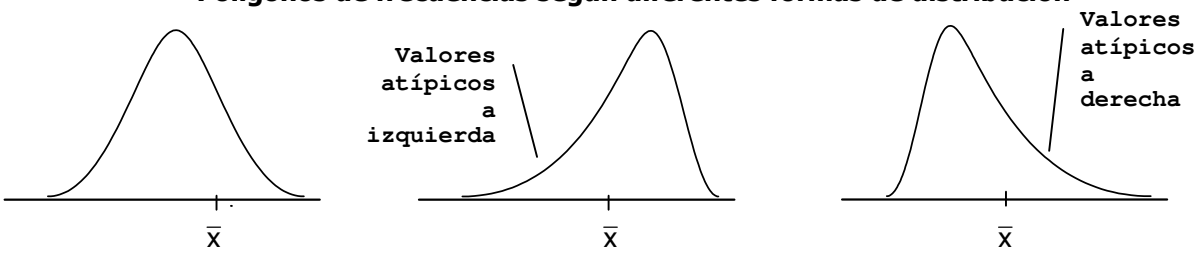
[5] Revisar el ejemplo del Párrafo Nº 2 de la Actividad Nº 1.

Cuando en la serie de observaciones existen valores extremos o *"atípicos"*, estos influirán en el valor de $\bar{x}$, pudiendo llegar a distorsionarlo de tal modo que no represente al *"común"* de los datos del conjunto (es una medida *"no resistente"*). Veamos el siguiente ejemplo:

$\bar{x} = 11,6$ es el promedio de los siguientes datos: $\{12, 10, 9, 16, 11\}$. En cambio, si el conjunto fuera $\{12, 10, 9, \mathbf{160}, 11\}$; el promedio resultaría: $\bar{x} = 40,4$. El valor atípico (160) afecta a $\bar{x}$ ***alejándola de la tendencia central*** del conjunto, resultando esta en un ***valor muy diferente*** al de los datos *normales* de la serie (12, 10, 9 y 11).

Entonces, ¿el promedio de 40,4 representa apropiadamente al *"común"* de los datos del conjunto? No, porque *"**no resiste**"* el efecto del valor extremo [6] y se ***desplaza de la tendencia central hacia el lado*** del valor atípico.

Polígonos de frecuencias según diferentes formas de distribución



Resumiendo: en un conjunto de datos en el cual los valores atípicos tienen un peso significativo (difieren mucho de los valores *"regulares"*), el ***promedio aritmético***, por ser una medida *"no resistente"*, ***debe ser analizado con cuidado.*** Esto es así porque -como en el ejemplo anterior- puede resultar fuertemente desplazado de la tendencia central e inducir a interpretaciones erróneas acerca del conjunto de datos que resume.

IMPORTANTE

La presencia de valores extremos en una distribución se manifiesta por formas (histogramas y polígonos de frecuencias) marcadamente asimétricas. De ahí la importancia de realizar una cuidadosa exploración previa (gráfica y numérica) de los datos.

- ***Tercera Propiedad***

Se denomina ***residuo o desvío individual*** de un dato cualquiera de la serie, con respecto ***a la media aritmética*** de todo el conjunto, a la ***diferencia entre el valor de ese dato y el valor*** de $\bar{x}$.

Retomando el ejemplo de las edades de los alumnos del curso de Estadística, el residuo o desvío con respecto a la edad promedio de 23 años, de cada uno de los datos del conjunto será:

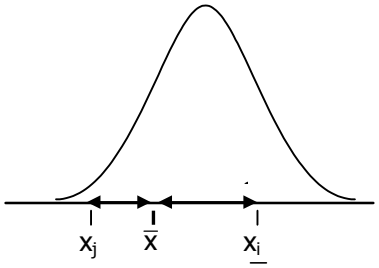
Dato (x_i)	**Desvío** ($d_i=x_i-23$)
19	-4
27	4
26	-3
28	5
....	...
x_i	x_i-23
...	...
30	7
	$\Sigma d_i=0$

[6] Los valores extremos pueden serlo por defecto o por exceso como en este ejemplo.

Cada desvío con respecto al valor de la media de todo el conjunto podrá ser negativo, nulo o positivo, según el valor del dato sea menor, igual o mayor al del promedio. Así, el desvío del primer dato x_1=19 años es: d_1=19-23=-4 años. El desvío del segundo dato x_2=27 años es: d_2=27-23=+4 años y así sucesivamente hasta el último dato x_{139}=30 años, cuyo desvío es: d_{139}=30-23=+7 años.

En forma simbólica, el desvío de un dato genérico x_i se expresa: $d_i = x_i - \bar{x}$ y para un conjunto $\{x_1, x_2, x_3, x_4, x_5, \ldots .. x_i, \ldots\ldots\ldots x_n\}$ de observaciones, habrá n residuos individuales $\{d_1, d_2, d_3, d_4, d_5, \ldots .. d_i, \ldots\ldots\ldots d_n\}$.

Es de notar que los desvíos (desprovistos del signo positivo o negativo) miden la ***"distancia"*** que separa a ***cada individuo*** observado del ***promedio general*** del grupo. Por ejemplo: el segundo individuo de la serie se diferencia en 4 años del promedio general de 23 años, mientras que la distancia al promedio del individuo 139, es de 7 años.



Los residuos de un conjunto de datos, con respecto a $\bar{x}$, tienen la propiedad de que la suma de todos ellos (cada uno con su signo negativo, nulo o positivo) es siempre igual a cero:

$$\sum_{i=1}^{n} \left(x_i - \bar{x} \right) = \sum_{i=1}^{n} di = 0$$

Es decir que, por esta propiedad, ***la suma*** (*-4+4-3+5.....+7*) de los 139 residuos individuales de las edades de los estudiantes de Estadística, **será igual a cero**[7].

- ***Cuarta Propiedad***

En ciertas ocasiones de trabajo disponemos de dos o más promedios aritméticos, que resumen a diferentes conjuntos de datos de ***una misma variable***.

Por ejemplo: por datos recogidos se sabe que el salario mensual promedio de n_1= 107 agentes públicos provinciales ***varones*** es $\bar{y} = \$1133,25$, mientras que el salario medio de n_2=73 empleadas ***mujeres*** es $\bar{z} = \$862,07$.

En estas condiciones podría resultar útil conocer el promedio que resume a los salarios de ***todos los agentes públicos***, considerados como un solo conjunto de observaciones (n = 180 en total). La *"**media de medias**"* es el promedio que resuelve cuestiones como la planteada. Esta *"media de medias"* a la que simbolizaremos con la notación $\bar{\bar{x}}$ (ó $\bar{\bar{z}}$ ó $\bar{\bar{y}}$), se define del siguiente modo:

Sea $\bar{y}$ la media aritmética de n_1 observaciones de cierta variable en estudio, y $\bar{z}$ la media de otro conjunto de n_2 datos de la misma variable; el promedio aritmético $\bar{\bar{x}}$ ***de ambas medias*** (*"media de medias")* será[8]:

$$\bar{\bar{x}} = \frac{n_1 \cdot \bar{y} + n_2 \cdot \bar{z}}{n_1 + n_2}$$

[7] Esta propiedad puede ser verificada en forma completa, utilizando el conjunto de 5 datos $\{12, 10, 9, 16, 11\}$ del ejemplo anterior.

[8] Es muy importante tener presente que los datos z_i e y_i deben ser *conceptualmente "promediables"* entre sí, de tal modo que $\bar{\bar{x}}$ represente un concepto válido y comprensible.

En consecuencia, *el salario promedio general de* **todos los agentes públicos** *del ejemplo será de $1023, 27 porque:*

$$\bar{x} = \frac{107 \,.\, 1133,25 + 73 \,.\, 862,07}{180} = \$1023,27$$

3.2. Cálculo de la Media

El procedimiento a seguir para el cálculo de $\bar{x}$ dependerá del estado en el que se encuentran los datos a trabajar. Esto es:

- ✓ ¿se trata de datos en el estado *"bruto"* de la matriz de datos (sin ninguna forma de resumen)?,
- ✓ ¿se trata de datos resumidos en un arreglo de frecuencias?,
- ✓ ¿se trata de datos resumidos en una distribución de frecuencias con intervalos?

IMPORTANTE

Recomendamos especialmente a los estudiantes del curso, familiarizarse con el manejo de algún *software* que les permita resolver los cálculos estadísticos mediante el uso de computadoras.

Seguidamente presentamos los procedimientos para el cálculo *manual* de $\bar{x}$ (con la ayuda de una calculadora común) con dos propósitos:

- que puedan revisar los conocimientos teóricos desde el cálculo aplicado a ejercicios concretos,
- que puedan resolver problemas de trabajo aun cuando no disponen del auxilio informático.

3.2.1. Datos sin resumir

El procedimiento de cálculo consiste en aplicar estrictamente y paso a paso, el concepto de la media aritmética. O sea: *"sumar todos los datos del conjunto y luego, dividir esa suma por el total n de observaciones de la serie"*.

3.2.2. Datos agrupados en arreglo de frecuencias

El resumen en arreglo de frecuencias permite identificar a cada dato por su valor individual y, por ello, el cálculo se realiza de igual modo que en la situación anterior*: sumando todas las observaciones individuales y dividiendo la suma por n*.

Retomemos el arreglo de frecuencias que resume la distribución de los alumnos del curso de Estadística, según las horas diarias que dedican a ver televisión.

Alumnos de Estadística según el tiempo diario que miran TV

Horas TV (x_i)	**Estudiantes (f_i)**
0	25
1	26
2	49
3	18
4	13
5	5
6	2
8	1
Total	**139**

El promedio de este grupo de datos será:

$$\bar{x} = \frac{\overbrace{0+0+....+0}^{25\ veces}+\overbrace{1+1+....+1}^{26\ veces}+\overbrace{2+2+....+2}^{49\ veces}+\overbrace{3+3+....+3}^{18\ veces}+.....+6+6+8}{139}$$

o sea:
$$\bar{x} = \frac{0.25+1.26+2.49+...+6.2+8.1}{139} = \frac{275}{139} = 2 \text{ horas diarias}$$

Es decir que, estando los datos resumidos en un arreglo de frecuencias, el procedimiento de cálculo de la media consiste en: *"multiplicar cada dato de la serie por su correspondiente frecuencia absoluta, sumar entre sí todos los productos y, finalmente, dividir la suma resultante por el total n de datos"*.

A esta forma de promediar los datos se la llama **"media ponderada por las frecuencias"** y simbólicamente se expresa como:

$$\bar{x} = \frac{\sum x_i . f_i}{n}$$

IMPORTANTE

Nótese que la media ponderada calculada a partir de un arreglo de frecuencias, reproduce ***estrictamente*** al ***concepto original*** del promedio, en tanto se trata de: *"la suma de todas las observaciones dividida por el total de datos*".

3.2.3. Datos agrupados en una distribución con intervalos

Cuando los datos se encuentran agrupados en una distribución con intervalos, es necesario basar el cálculo de $\bar{x}$ en un procedimiento que no considere a los valores individuales, ya que estos no son conocidos en esta situación de trabajo.

En el ejemplo siguiente se presenta la distribución de n = 72 "grupos turísticos"[9] observados en Puerto Iguazú, resumidos en intervalos del "*gasto total*[10] *del grupo en un día completo de estadía*" en el lugar.

Turistas Según Gasto de un Día -Pto. Iguazú. Febrero'94-

Gasto ($)	Grupos (f_i)	Pto. Medio (x_i)
00 - 55	19	27,5
55 - 110	20	82,5
110 - 165	18	137,5
165 - 220	7	192,5
220 - 275	4	247,5
275 - 330	3	302,5
330 - 385	1	357,5
Total	**72**	

Fuente: "ESTUR 93/94". CFI-FHyCS (UNaM)

La tabla permite saber, por ejemplo, que 20 grupos gastaron en un día entre $55 y $110, pero no es posible conocer el gasto exacto de cada uno de ellos individualmente.

[9] Conjunto de personas (familiares o no) que comparten el mismo presupuesto de viaje.

[10] Comprende el gasto por todo concepto (alojamiento, alimentación, transporte, esparcimiento, servicios varios, compras, etc.) por *"grupo turístico"*, en 24 horas corridas de permanencia en Pto. Iguazú.

El cálculo de la media en esta situación de trabajo, se basa en asumir a cada dato individual (desconocido) como **equivalente al valor del punto medio** o "marca" de la clase en que se ubica. Por ejemplo, se asumirá que el gasto individual de cada uno de los 18 grupos comprendidos entre $110 y $165, fue equivalente a $137,5. De igual modo, asumiremos que el gasto individual de cada grupo comprendido entre $275 y $330 fue equivalente a $302,5 y así sucesivamente para todos los datos de la distribución.

Al reemplazar los datos individuales por el valor del punto medio de clase que los representa, el promedio resultará de un cálculo similar al anterior. Es decir:

$$\bar{x} = \frac{27{,}5 \,.\, 19 + 82{,}5 \,.\, 20 + 137{,}5 \,.\, 18 + 192{,}5 \,.\, 7 + 247{,}5 \,.\, 4 + 302{,}05 \,.\, 3 + 357{,}5 \,.\, 1}{72}$$

O sea:

$$\bar{x} = \frac{8085}{72} = \$112{,}30 \text{ de gasto promedio diario por grupo}$$

Nuevamente, la media se obtiene por un procedimiento "ponderado por las frecuencias" del tipo $\bar{x} = \frac{\sum x_i . f_i}{n}$, en el cual los valores "x_i" ahora son las **marcas de cada clase** y los valores "f_i" son las correspondientes **frecuencias absolutas de clase**.

IMPORTANTE

Nótese que el valor de la media que resulta por esta forma de cálculo no es exacto, en tanto se basa en los puntos medios de clase y no en los datos originales. Se obtiene entonces, un valor "aproximado" al "verdadero valor" del promedio.

Actividad Nº 2

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 2 de la Guía de Actividades correspondiente a esta unidad **.**

4. La Mediana

A diferencia de los promedios (la media aritmética en nuestro caso) que resultan de una operación **basada en todos los datos** de la serie, la mediana marca la tendencia central del conjunto tomando en consideración a **uno solo de ellos**.

Concepto

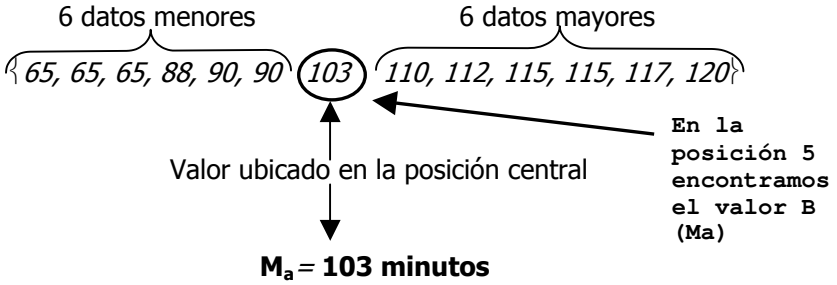
La mediana ($\mathbf{M_a}$) de una distribución es **el dato** que ocupa **la posición central** del conjunto de observaciones, debiendo estar los datos **previamente ordenados** en forma ascendente (o descendente) de magnitud.

Símbología: son diversos los símbolos aceptados para representar a esta medida: M_{dn}, M, M_{ed}, M_d, M_e, X_5, X_{me}; entre otros. Nuevamente, las letras mayúsculas y minúsculas se reservan para distinguir lo "*poblacional*" de lo "*muestral*". En este curso emplearemos indistintamente la notación $\mathbf{M_a}$.

Consideremos como ejemplo la siguiente serie de datos numéricos, referidos al "*tiempo en minutos*" que le requirió realizar un examen de Estadística a un grupo de n = 13 alumnos:

Minutos: { 120, 65, 110, 117, 65, 115, 88, 90, 103, 112, 90, 65, 115 }

El conjunto **ordenado** *en forma ascendente*[11] *resulta:*



M_a = 103 minutos

Es decir que la mediana es el **valor que se ubica en el centro del conjunto de datos ordenados** y, como tal, divide a la serie en **dos grupos con igual cantidad de observaciones** (aproximadamente la mitad): uno de ellos contiene a todos los **casos que son inferiores o iguales** al valor mediana, y el otro a todos los **casos iguales o superiores** *a él.*

Por ello, la M_a representa al **"individuo medio"** de la muestra o población en estudio: (en esta característica observada) el alumno que utilizó **"103 minutos"** para resolver el examen, es el alumno medio del grupo, ya que por debajo de él se ubican la mitad de sus compañeros y por encima la otra mitad.

4.1. Principales Propiedades de M_a

- **Primera Propiedad**

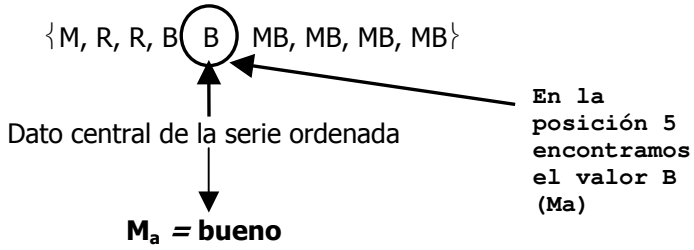
  Es una medida basada en un concepto fácilmente comprensible, que requiere de operaciones simples para aplicarla (ordenar y ubicar la posición central).

- **Segunda Propiedad**

  Siendo M_a el dato que ocupa el ***lugar central de la distribución ordenada***, el concepto tiene significado y, en consecuencia, ***es aplicable a datos categóricos ordinales***. Veamos el ejemplo siguiente en el que se analizan las respuestas sobre la *"calificación a la Fiesta Provincial de La Flor*[12]*"* (Montecarlo, Misiones, año 2001), obtenidas en un relevamiento efectuado a $n = 9$ personas mayores de 16 años que asistieron al evento.

  Calificaciones: {R, MB, MB, B, M, MB, R, MB, B}

  El conjunto **ordenado** resultará:



M_a = bueno

A ambos lados de la categoría mediana se ubica la misma cantidad de observaciones, unas de **categoría igual o inferior** a M_a y las otras, de **categoría igual o superior** a ella.

Es decir, aproximadamente *el 50% de los visitantes del ejemplo, asignó a la Fiesta una calificación "buena" o inferior y la otra mitad la calificó como "buena" o superior.*

- **Tercera Propiedad**

  La mediana de ***datos numéricos*** tiene la propiedad de ser ***"resistente"*** a la presencia de valores extremos en el conjunto de observaciones. Retomando el ejemplo de los minutos que les

[11] Idéntico resultado se obtendría si el orden en los datos fuera descendente.

[12] Las categorías posibles de respuesta fueron: muy bueno ***(MB)***, bueno ***(B)***, regular ***(R)***, malo ***(M)*** y muy malo ***(MM)***.

llevó a los 13 alumnos de Estadística realizar el examen, si ***reemplazáramos*** el dato del primer alumno (65) por el valor 5 minutos; la ***mediana del conjunto permanecería inalterada*** en:

$$M_a = 103 \text{ minutos}$$

Lo mismo ocurriría si se **reemplazara** el dato más alto de la serie (120) por cualquier valor atípico para ese conjunto de observaciones (por ejemplo 720 ó 7200).

Nótese que en estos ejemplos, la cantidad de n = 13 observaciones de la serie se mantiene inalterada, ya que suponemos la sustitución de un valor original por otro atípico. Es decir, la M_a **es resistente a valores extremos si no se modifica el tamaño n** del conjunto de datos.

- **Cuarta Propiedad**

En cambio, si al conjunto original se agregaran 2 nuevos alumnos (ahora $n = 15$) con 109 y 118 minutos respectivamente, la serie ordenada resultaría:

$$\{65, 65, 65, 88, 90, 90, 103, \underset{\uparrow}{109}, 110, 112, 115, 115, 117, 118, 120\}$$

$$\mathbf{M_a = 109 \text{ minutos}}$$

Es decir que la M_a es una medida que puede alterarse si se ***modifica la cantidad de datos*** de la serie.

- **Quinta Propiedad**

Por ser una medida que representa a todo el conjunto de datos mediante uno solo de sus valores, ***cuando se trabaja con datos numéricos*** la M_a no aporta elementos sobre la conformación general del grupo de observaciones (e individuos en consecuencia): *¿hay datos atípicos en la distribución?*, *¿cuán diferentes son los valores extremos en relación con los datos "comunes"?*

Retomando el ejemplo de Actividad Nº 2, si dijéramos que: *"la mitad de los 97 funcionarios (incluidos los 7 cargos gerenciales) de la empresa perciben haberes netos mensuales superiores a $753"*[13]; sin conocer los datos originales, no sabríamos que en el conjunto en estudio se incluyen valores tan extremos como $4927,....., $5124,...$6701 y $6890.

4.2. Determinación de la M_a

El procedimiento a seguir para determinar[14] el valor mediana de una distribución en estudio, dependerá del tipo de datos que se trate (numéricos u ordinales) y del estado de elaboración en que se encuentran (datos brutos, arreglos de frecuencias, distribución con intervalos).

4.2.1. Datos numéricos sin resumir

- Si el número de observaciones es impar

Cuando los datos en análisis son **numéricos** y el **número n** de observaciones que forman el conjunto **es impar**, habrá un **único valor** que ocupará la **posición central** del conjunto ordenado (ejemplos anteriores de n = 13 ó n = 15 estudiantes en el examen de Estadística). En esta situación el procedimiento consistirá en **ordenar rigurosamente** los datos por su magnitud (sentido ascendente o descendente) y luego, **identificar el valor que se ubica en el lugar central del conjunto ordenado** (que deja igual cantidad de datos a ambos lados). Ese valor es la mediana del conjunto.

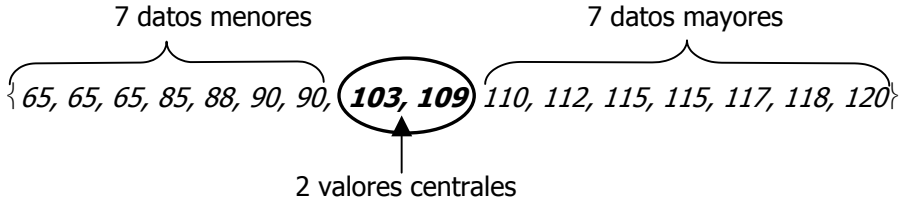
- Si el número de observaciones es par

Cuando el **número n** de observaciones de la serie **es par**, serán dos los valores centrales del conjunto ordenado, que separarán la misma cantidad de datos hacia

[13] Recomendamos realizar el ejercicio de verificar la exactitud de esta afirmación.

[14] Nótese que hablamos de *"determinar"* y no de *"calcular"* M_a, porque se trata de una medida "no calculada". Si bien analizaremos procedimientos basados en fórmulas y cálculos numéricos con los datos, en todos los casos se trata de razonamientos para **identificar el valor central** de la serie ordenada, tal como se define esta medida.

ambos lados. Por ejemplo, supongamos que fueron n = 16 los alumnos que rindieron el examen de Estadística:



En este caso la M_a se determina **por convención**, **promediando ambos datos centrales**. Es decir:

$$Ma = \frac{103+109}{2} = 106 \text{ minutos}$$ [15]

4.2.2. Datos numéricos en arreglo de frecuencias

En esta situación de trabajo el razonamiento debe seguir los mismos pasos anteriores, considerando que en el arreglo de frecuencias los **datos ya se encuentran ordenados por magnitud**. El problema entonces consiste en:

a- ubicar el lugar central del conjunto ordenado (posición del valor M_a),

b- identificar el valor (o los valores si n es par) que ocupa esa posición (o esas posiciones).

Retomemos como ejemplo la distribución de los alumnos del curso de Estadística, según las horas diarias que dedican a la TV:

Alumnos de Estadística según el tiempo diario que miran TV

Horas TV (x_i)	**Estudiantes (f_i)**	F_a
0	25	25
1	26	51
2 ← ***M_a***	**49**	**100**
3	18	118
4	13	131
5	5	136
6	2	138
8	1	139
Total	**139**	

a- Ubicación del lugar central de la distribución ordenada

- Si el número de observaciones es impar (ej.: n = 139), el conjunto ordenado de menor a mayor ocupará 139 posiciones [16] y una sola de ellas será la central: $\frac{139+1}{2} = 70$, de tal modo que a su izquierda quedarán 69 datos menores o iguales y a su derecha otros 69 datos mayores o iguales.

Tenemos así que, tratándose de un **número impar** de observaciones, la **posición o lugar central** de la distribución se determina mediante:

[15] Notar que en este caso, M_a no es exactamente un dato de la serie. La medida toma el valor "teórico" que resulta de promediar los dos datos centrales y, en consecuencia, ocupa un lugar también "teórico", ubicado entre ambos valores.

[16] Imagine a los 139 valores individuales ordenados uno al lado del otro sobre una recta horizontal. El primero será "0" (se repite por 25 veces) y el último será 8 (una sola vez).

$$\text{Posición Ma} = \frac{n+1}{2}$$

- Si el número de observaciones es par (ej.: $n = 160$ alumnos), serán dos las posiciones centrales ($\text{Posición } 80 = \frac{160}{2}$ y $\text{Posición } 81 = \frac{160}{2} + 1$) las que dejan igual cantidad de observaciones hacia ambos lados (79 en este caso).

Tratándose de un **número par de datos**, las **dos posiciones centrales se determinan** mediante:

$$\text{Posición}_1 = \frac{n}{2} \quad y \quad \text{Posición}_2 = \frac{n}{2} + 1$$

b- Determinación del valor Ma

Habiendo identificado la posición central (o las dos posiciones cuando n es par) del conjunto ordenado, el problema ahora es identificar el dato (o los datos) que se ubica(n) en ese lugar. Para ello nos valemos de las frecuencias acumuladas (en el sentido "menor que"), razonando en el ejemplo anterior del siguiente modo:

- ✓ Hasta el valor 1 de la distribución **se acumulan** 51 datos ordenados y, en consecuencia, ninguno de ellos (valores 0 y 1 del arreglo) alcanzan la **posición 70**.
- ✓ Al pasar al valor 2 ya son 100 las observaciones acumuladas, lo que significa que uno de los 49 datos iguales a 2 es el que ocupa la posición central 70.
- ✓ Es decir: la $M_a = 2$ horas diarias.

Este valor de la mediana nos indica que "aproximadamente la mitad de los alumnos entrevistados dedica 2 horas diarias o menos a ver TV" (obviamente la otra mitad, dedica 2 horas o más por día).

El razonamiento es idéntico cuando el **número n de casos** del conjunto **es par**, teniendo en cuenta que ahora el problema consiste en identificar los valores que ocupan las **dos posiciones centrales** y luego, determinar M_a como el promedio entre ambos datos.

4.2.3. Datos *numéricos* en una distribución con intervalos

En esta situación de trabajo la mediana no puede ser determinada **exactamente** porque, al ser **desconocidos** los **datos individuales** que forman el conjunto en estudio, no hay manera de reconocer el valor que ocupa la posición central de la serie ordenada [17]. Por ello, el procedimiento consiste en **estimar** la M_a mediante el siguiente razonamiento:

a. determinar el punto medio "teórico" (o centro geométrico) de la serie haciendo:

$$\text{Posición Ma} = \frac{n}{2}$$

b. analizando las frecuencias acumuladas ("menor que"), identificar la clase o intervalo ("clase mediana") de la distribución en la que se ubica dicha posición;

c. **estimar** el valor mediana aplicando la siguiente **fórmula de interpolación**:

$$Ma = L_i + \frac{\frac{n}{2} - F_{a_{(i-1)}}}{f_i} \cdot a$$

siendo:

M_a : valor estimado de la mediana,
L_i : límite inferior de la "clase mediana",
$\frac{n}{2}$: punto medio de la serie de datos,
$Fa_{(i-1)}$: frecuencia acumulada anterior a la "clase mediana",
f_i : frecuencia absoluta de la "clase mediana",
a : amplitud de la "clase mediana".

Retomemos el ejemplo del gasto diario de los turistas en Pto. Iguazú

[17] Es de notar que los datos se encuentran ***ordenados por la magnitud de sus intervalos.***

Turistas según Gasto de un Día -Pto. Iguazú. Febrero'94-

Gasto ($)	Grupos (f_i)	F_a
00 - 55	19	19
55 - 110	**20**	**39**
110 - 165	18	57
165 - 220	7	64
220 - 275	4	68
275 - 330	3	71
330 - 385	1	72
Total	**72**	

clase M_a →

Fuente: "ESTUR 93/94". CFI-FHyCS (UNaM)

- Punto medio de la distribución:

$$\frac{n}{2} = \frac{72}{2} = 36$$

- Analizando las frecuencias acumuladas se observa que la primera clase reúne a los **primeros 19 datos** ordenados de la distribución y, en consecuencia, ninguno de ellos alcanza al **punto medio 36**.

  Al pasar a la segunda clase ya **son 39 los datos acumulados** en sentido ascendente de magnitud, razón por la cual entre los 20 datos de esta clase se encuentran los dos valores centrales de la distribución. Es decir, ésta es la "clase mediana"[18].

- Localizada la clase donde se ubica M_a, su **valor estimado** resultará de hacer:

$$Ma = 55 + \frac{36 - 19}{20}.55 = \$101,75$$

lo que permite decir: "la mitad de los grupos turísticos tienen un gasto diario de aproximadamente $101,75 o menos".

4.2.4. Datos categóricos ordinales

Cuando los datos en análisis son **ordinales** y se encuentran resumidos en una tabla de frecuencias, el procedimiento sigue un razonamiento similar al de la situación "datos numéricos en arreglo de frecuencias". O sea:

a- ubicar el lugar central (o los lugares si n es par) del conjunto ordenado (posición de la categoría M_a),

b- identificar el valor (o los valores si n es par) que ocupa esa posición (o esas posiciones).

Consideremos el ejemplo sobre los usuarios de la empresa misionera de servicios eléctricos:

Opinión de los Usuarios sobre el Servicio Eléctrico de Mnes. (EMSA)

Opinión	Usuarios	F_a
Muy Malo	3	3
Malo	20	23
Regular	151	174
Bueno	**469**	**511**
M. Bueno	42	685
TOTAL	**685**	

M_a →

Fuente: Departamento TISE-FHyCS. 1994

[18] La clase de la mediana siempre es aquella cuya frecuencia acumulada "menor que", resulta ***igual o inmediatamente mayor*** a: $\frac{n}{2}$ ó $\frac{n+1}{2}$, según corresponda.

- Posición central (en este caso *n es impar*): $\frac{n+1}{2} = \frac{686}{2} = 343$
- **Localizada la posición central** del conjunto ordenado, nos valemos de las frecuencias acumuladas para **identificar al dato que se ubica en ese lugar**. La categoría "muy malo" acumula 3 observaciones, la categoría "malo", 23 observaciones y 174 son las opiniones "regular" o menos. Al pasar a la categoría siguiente ya son 511 los datos acumulados, razón por la que uno de los 469 datos "bueno" es el que ocupa el lugar central 343. En consecuencia M_a = "bueno".

Así: "*aproximadamente la mitad de los usuarios entrevistados, tienen una opinión* ***"buena" o superior*** *sobre el servicio eléctrico que reciben*".

Si el número *n* de datos de la serie fuera **par** (por ejemplo n = 734 usuarios), existirían dos posiciones centrales: $\text{Posición}_1 = \frac{n}{2}$ y $\text{Posición}_2 = \frac{n}{2} + 1$ (lugares 367 y 368 en nuestro ejemplo). Con la ayuda de las frecuencias acumuladas, se podrá localizar la $\boldsymbol{M_a}$ identificando los datos (categoría) que se ubican en estos lugares.

Actividad Nº 3

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 3 de la Guía de Actividades correspondiente a esta unidad***.**

5. El Modo

Concepto

El modo ($\mathbf{M_o}$) de un conjunto de observaciones (numéricas o categóricas nominales u ordinales) es el **dato individual que más veces se repite** en la serie. $\mathbf{M_o}$ será el valor más típico, más recurrente o bien, el que reúne la **mayor frecuencia absoluta** entre todos los valores (categorías) individuales observados en el conjunto de datos que se analiza.

Símbología: algunos símbolos utilizados para representar a esta medida son: M_{do}, X_{mo}; entre otros. Nuevamente, las letras mayúsculas y minúsculas se reservan para distinguir lo "*poblacional*" de lo "*muestral*". En este curso emplearemos indistintamente la notación $\mathbf{M_o}$.

En este caso tenemos también una medida que toma en consideración a **una sola de las observaciones**, aunque no siempre se ubica en los valores centrales de la serie de datos.

Tomando como ejemplo la serie de datos referidos al "tiempo en minutos" que le requirió realizar un examen a los alumnos de Estadística y a las "calificaciones a la Fiesta Provincial de La Flor", realizadas por 9 visitantes al evento, tendremos:

dato más frecuente

Minutos: { *65, 65, 65, 88, 90, 90, 103, 110, 112, 115, 115, 117, 120* }

M_o= 65 minutos

dato más frecuente

Calificaciones: { M, R, R, B, B, MB, MB, MB, MB }

M_o = muy bueno

5.1. Principales Propiedades del M_o

- **Primera Propiedad**

  Es una medida conceptualmente simple, fácil de interpretar y de comunicar, que requiere únicamente del conteo para ser determinada.

- **Segunda Propiedad**

  Por no requerir de ninguna forma de orden en los datos, tiene significado y ***es aplicable a datos categóricos nominales*** (es la única de las tres medidas de tendencia central que hemos tratado, posible de ser utilizada con este tipo de datos).

- **Tercera Propiedad**

  Cuando la diferencia entre la frecuencia máxima observada (frecuencia modal) con alguna de las restantes no es muy grande, el M_o como medida característica de la distribución pierde relevancia.

IMPORTANTE

Puede ocurrir que en un conjunto de datos se encuentren dos o más valores que reúnen la misma frecuencia absoluta máxima[19] (en nuestros ejemplos si tuviéramos ***dos alumnos más*** con 90 y 115 minutos respectivamente o bien, ***dos visitantes más*** que califiquen la *Fiesta de la Flor* como *Regular*). En tales casos las distribuciones resultarían ***bimodal*** (dos valores con la misma frecuencia máxima) o ***multimodal*** (tres o más valores con esta propiedad) y no es posible determinar un único valor/categoría M_o para toda la serie.

5.2. Determinación del M_o

5.2.1. Para arreglos de frecuencias y datos categóricos

Si los datos individuales se encuentran sin agrupar, lo recomendable es resumirlos previamente en un arreglo de frecuencias (o en una tabla de frecuencias para datos categóricos). Encontrándose los datos presentados de esta manera, la determinación del M_o simplemente se remite a ubicar en la distribución, el valor o categoría al que corresponde la mayor frecuencia absoluta.

Consideremos el siguiente ejemplo:

Estudiantes del Curso de Estadística. FHyCS-Año 2001

según Sexo

Sexo	Estudiantes
Varón	30
Mujer ← M_o	**109**
Total	**139**

Fuente: elaboración propia.

según el Tiempo Diario que Miran TV

Horas TV (x_i)	Estudiantes (f_i)
0	25
1	26
2 ← M_o	**49**
3	18
4	13
5	5
6	2
8	1
Total	**139**

[19] Esta situación es muy raro que ocurra si el número (n) de observaciones es "suficientemente grande".

Usuarios del Servicio Eléctrico de Misiones (EMSA), Según Opiniones sobre la Calidad del Servicio

Opinión	Usuarios
M. Bueno	42
Bueno	**469**
Regular	151
Malo	20
Muy Malo	3
TOTAL	**685**

M_o

Fuente: Departamento TISE-FHyCS. 1994

Así entonces:

"*las mujeres predominan en el grupo de estudiantes de Estadística y lo más común o frecuente son los alumnos que dedican 2 horas diarias a ver TV*", *y*
"*la opinión de que el servicio eléctrico es bueno, es la más típica entre los usuarios de la Empresa de Electricidad de Misiones*".

5.2.2. Para una distribución con intervalos

En la situación de trabajo en la que los datos son numéricos y se encuentran resumidos en una distribución con intervalos (como el ejemplo de los gastos turísticos que se presentan a continuación), el M_o debe determinarse mediante el siguiente **procedimiento de estimación**, aceptado por convención:

Turistas según Gasto de un Día -Pto. Iguazú. Febrero'94-

Gasto ($)	Grupos (f_i)
00 - 55	19
55 - 110	**20**
110 - 165	18
165 - 220	7
220 - 275	4
275 - 330	3
330 - 385	1
Total	**72**

clase modal

Fuente: "ESTUR 93/94". CFI-FHyCS (UNaM)

Asumiendo que la clase que presenta la **mayor frecuencia absoluta** de la distribución (**"clase modal"**) es la que **contiene entre sus datos** al valor modal, una vez identificada el valor del M_o se puede **estimar** mediante el siguie*nte* **procedimiento de interpolación**:

$$M_o = L_i + \frac{d_1}{d_1 + d_2} . a$$

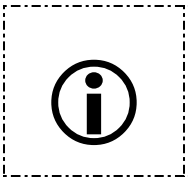
siendo:
L_i: límite inferior de la clase modal,
d_1: la diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase inmediata anterior a la modal,
d_2: la diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase inmediatamente posterior a la modal,
a: amplitud de la clase modal.

En nuestro ejemplo resultará:

L_i = 55 **d_1 =** 20-19 = 1 **d_2 =** 20-18 = 2 **a =** 55

$$Mo = 55 + \frac{1}{1+2} . 55 = \$73,3 \text{ diarios}$$

O sea: "*estimamos que el gasto más frecuente entre los 72 casos observados, es de $73,3 diarios*".



IMPORTANTE

Este procedimiento para estimar el modo de datos numéricos agrupados en clases es ***altamente sensible a la forma*** en que se define la distribución. Esto es: al número de intervalos y a la amplitud de cada uno de ellos.

El siguiente ejemplo ilustra sobre este problema. El mismo grupo de n = 9 datos se organiza de 3 maneras distintas:

Situación A

Datos	fi
65 ← M_o	**2**
70	1
72	1
73	1
81	1
82	1
86	1
87	1
Total	**9**

Situación B

Datos	fi
65 - 69	2
70 – 74 ← *Clase M_o*	**3**
75 – 79	0
80 – 84	2
85 - 89	2
Total	**9**

Situación C

Datos	fi
65 - 69	2
70 – 79 ← *Clase M_o*	3
80 - 89	**4**
Total	**9**

El **modo verdadero** de la serie es $M_o = 65$ ya que se trata del valor del conjunto con mayor frecuencia (Situación A).

En la Situación B la **clase modal** es la segunda de la distribución (**70–74**) y aplicando el procedimiento de estimación por interpolación resulta: $M_o = 70{,}75$.

En la Situación C el M_o se ubicará en la tercera clase (**80–89**), resultando su estimación: $M_o = 81{,}5$ [20].

Actividad Nº 4

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 4 de la Guía de Actividades correspondiente a esta unidad.

6. Cuartiles, Deciles, Centiles

Utilizando medidas de tendencia central podemos describir a los grupos turísticos observados en Puerto Iguazú señalando, por ejemplo, que:

"*se trata de grupos que observan un promedio de $112,30 diarios de gasto por todo concepto; siendo $73,30 la suma que diariamente gastan con mayor frecuencia y la mitad de los grupos analizados destina $101,75 o más por día a satisfacer sus necesidades*".

Esta descripción permite una buena **comprensión global** de los datos elaborados y, por ende, de los individuos analizados; pero muy poco o nada nos informa sobre aspectos más específicos del fenómeno en estudio. Por ejemplo:

- ✓ ¿por encima de qué valor se ubican los turistas que más gastan? o en términos más concretos, ¿qué nivel del gasto corresponde al 10% de los turistas que más gastan?,
- ✓ ¿por debajo de qué monto se ubican los grupos que menos gastan diariamente?,
- ✓ ¿entre qué valores están los niveles de gastos centrales?,
- ✓ etc.

[20] Sugerimos verificar los resultados de las situaciones A y B.

Es decir, en la descripción de un conjunto de datos, **las medidas de tendencia central no dan cuenta de la diversidad de situaciones (variabilidad o dispersión) que se presentan**. Es preciso entonces, agregar a esta información otros elementos que permitan una descripción más completa, haciendo referencia a otras características de la distribución [21].

En todo conjunto de datos (numéricos u ordinales) se pueden determinar **ciertos valores característicos que amplían la información** proporcionada por las medidas sintéticas de tendencia central sobre los individuos que se analizan. Estos datos, ubicados en posiciones estratégicas del conjunto, permiten conocer aspectos de su composición y estructura, que aportan nuevos elementos para el análisis. Es decir, las preguntas señaladas precedentemente, pueden responderse a partir de ciertos **datos ubicados estratégicamente** en una **distribución ordenada**.

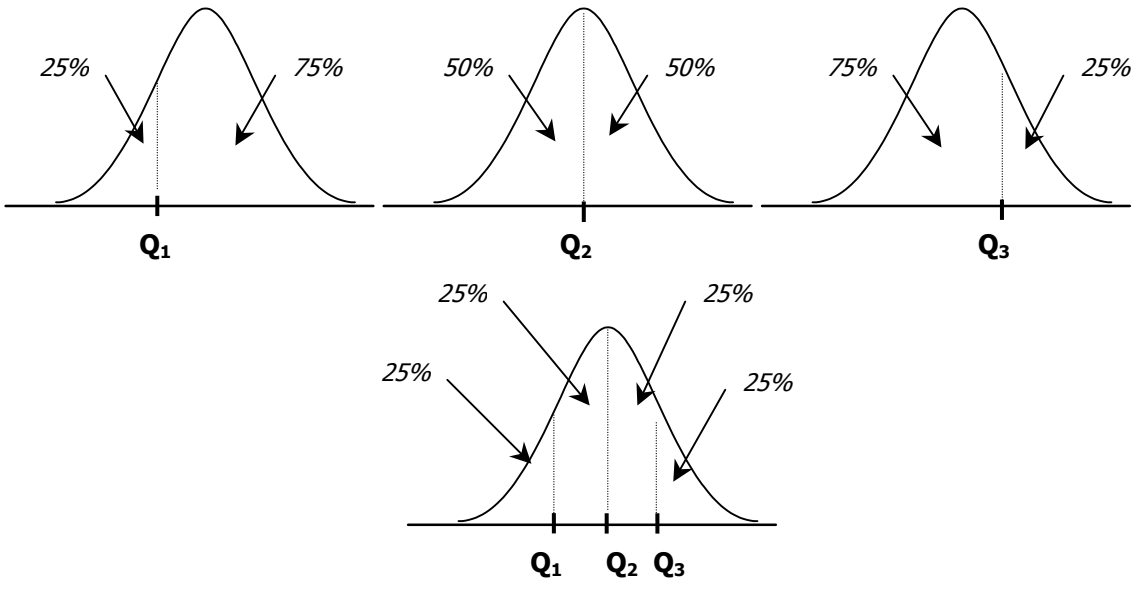
Las *medidas de posición*: **cuartiles, deciles y centiles**, son las que permiten individualizar a los datos que reúnen las condiciones señaladas.

6.1. Los Cuartiles

En toda distribución de datos **numéricos o categóricos ordinales** es posible hallar **tres observaciones individuales** que dividen al **conjunto, previamente ordenado en forma ascendente**, en cuatro partes iguales, cada una de ellas con el 25% de los datos.

- **Cuartil 1 – Primer Cuartil (Q_1):** es aquel valor del conjunto de observaciones que se ubica en una posición tal que a uno de sus lados deja una cuarta parte (25%) de los datos que son menores o iguales a él, y hacia el otro lado las tres cuartas partes (75%) de los datos que son mayores o iguales que él (es el valor que se ubica en la posición ¼ del conjunto ordenado).
- **Cuartil 2 – Segundo Cuartil (Q_2):** coincide con la mediana ya que divide al conjunto en dos partes, cada una con la mitad de los datos: $Q_2 = M_a$.
- **Cuartil 3 – Tercer Cuartil (Q_3):** es el dato situado en la posición que deja ¾ de las observaciones menores o iguales que él hacia un lado y ¼ de las observaciones mayores o iguales que él hacia el otro lado (el dato que se ubica en la posición ¾ de la serie ordenada).

Gráficamente



[21] Una vez más: no se trata de reducir la descripción de un conjunto de datos en un único valor, por más expresivo que el mismo pueda resultar, sino de comunicar *la forma* de la distribución en la que se expresa la disparidad y repetición de los valores de la variable.

Ejemplo:

Para la distribución de los grupos turísticos según el nivel de gasto diario en Iguazú, los cuartiles resultan:

Q_1 = \$52,11 Q_2 = Ma = \$101,75 Q_3 = \$155,83

Es decir que:

"*Una cuarta parte de los grupos (los 18 grupos que menos gastan) registra un nivel de gasto diario igual o inferior a \$52,11, mientras que el 25% de los que más gastan se ubican en \$155,83 ó más por día. Es decir que el 50% (36) de los grupos centrales registra un nivel de gasto comprendido entre \$52,11 y \$155,83 diarios*".

"*Considerando que el gasto mediana es de \$101,75, una cuarta parte de los turistas registra gastos diarios entre \$52,11 y \$101,75, y otra cuarta parte gasta entre \$101,75 y \$155,83*".

Determinación de los Cuartiles

El procedimiento para determinar Q_1 y Q_3 de una distribución sigue un razonamiento análogo al de la mediana, pero considerando que ahora se trata de identificar a los datos localizados en las posiciones ¼ y ¾ del conjunto ordenado. Para ello procedemos de la siguiente manera:

- **Localizamos las posiciones de los cuartiles**; la manera más sencilla de obtenerlas es:

$$\text{Posición } Q_1 = \frac{n}{4} \quad \text{y} \quad \text{Posición } Q_3 = 3.\frac{n}{4}$$

En nuestro ejemplo de los gastos turísticos, la posición del cuartil 1 será:

$$\text{Posición } Q_1 = \frac{72}{4} = 18$$

- Posteriormente, **inspeccionando las frecuencias acumuladas**, individualizamos los datos que ocupan las posiciones cuartílicas deseadas.
- **Cuando los datos son numéricos y se encuentran resumidos en una distribución con *intervalos***, primero debemos ***ubicar la clase del cuartil***, y luego ***estim*ar** su valor mediante el siguiente cálculo:

$$Q_1 = L_i + \frac{\frac{n}{4} - Fa_{(i-1)}}{f_i}.a \qquad \text{y} \qquad Q_1 = L_i + \frac{\frac{3.n}{4} - Fa_{(i-1)}}{f_i}.a$$

Donde los datos a considerar en cada una de estas expresiones (L_i, $Fa_{(i-1)}$, f_i, a) toman como referencia a las clases de Q_1 y Q_3 respectivamente, con significado idéntico al explicado para determinar la M_a en esta situación de trabajo.

En el ejemplo de los gastos turísticos:

La clase del cuartil 1 es la primera (0-55), por consiguiente podemos estimar el Q_1 de la siguiente manera:

$$Q_1 = L_i + \frac{\frac{n}{4} - Fa_{(i-1)}}{f_i}.a = 0 + \frac{18-0}{19}.55 = 52,11$$

Siguiendo el procedimiento indicado, verifique el valor correspondiente al tercer cuartil.

6.2. Los Deciles

Son los nueve valores de la distribución ordenada en forma ascendente que la dividen en diez partes iguales, cada una de ellas con el 10% de los datos.

- **Decil 1 – Primer Decil (D_1):** es aquel *valor del conjunto de observaciones* que se ubica en una posición tal que, a uno de sus lados *deja al 10% de los datos que son menores o iguales a él* y,

hacia el otro lado, el 90% de los datos restantes que son mayores o iguales que él (es el valor que separa el primer décimo del conjunto ordenado en forma ascendente).

- **Deciles 2, 3, 4, 5, 6, 7, 8 y 9 (D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9):** se definen trasladando el concepto de D_1 al segundo décimo, tercer décimo....., noveno décimo de la serie ordenada en forma ascendente ($D_5 = M_a$).

  En este caso, la forma sencilla de **ubicar la posición** de un decil genérico "i" (para i = 1, 2, 3, 4, 5, 6, 7, 8 ó 9) será mediante el cociente:

$$\frac{i \,.\, n}{10}$$

Luego, la determinación seguirá los pasos ya explicados y la **estimación por interpolación** se basará en:

$$D_i = L_i + \frac{\frac{i.n}{10} - Fa_{(i-1)}}{f_i} \,.\, a$$

6.3. Los Centiles (C_1, C_2, ..., C_{98}, C_{99})

Son noventa y nueve valores de la distribución ordenada en forma ascendente, que la dividen en cien partes iguales, cada una de ellas con el 1% de los datos.

La posición del "i"-ésimo centil (siendo i = 1, 2, 3,....., 98 ó 99) se determina por:

$$\frac{i.n}{100}$$

La estimación por interpolación resulta de aplicar la siguiente operación **a la clase del centil** genérico "i":

$$C_i = L_i + \frac{\frac{i.n}{100} - Fa_{(i-1)}}{f_i} \,.\, a$$

	***Actividad Nº 5*** *Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 5 de la Guía de Actividades correspondiente a esta unidad***.**

6.4. La curva de Lorenz asociada a medidas de posición

Como vimos en la **unidad anterior**, es posible asociar a cierto tipo de variables (ingreso, propiedad de la tierra, etc.) la gráfica de Lorenz, que nos permitirá analizar el grado de concentración/distribución de estos recursos en la población en estudio. En aquel momento, se presentó la construcción de esta gráfica a partir de una **tabla de frecuencias construida en base a intervalos de igual amplitud**; sin embargo, es posible hacerlo construyendo intervalos de distinta amplitud, cada uno de los cuales incluya la misma cantidad de individuos, de tal forma que la frecuencia relativa porcentual en cada uno de ellos sea del 25%, o del 10%, etc. Esto significa construir intervalos cuyo límite superior coincide con los cuartiles (tendríamos cuatro intervalos), o con los deciles (diez intervalos), etc.

Consideremos por ejemplo la distribución de los hogares según el ingreso familiar en la ciudad de Formosa. Se puede ver en el Cuadro siguiente que los hogares aparecen distribuidos en intervalos de clase de diferente amplitud, de manera que cada uno de los mismos agrupa aproximadamente un 10% del total de los hogares (4329 hogares). De esta manera estamos presentando los datos en una **distribución según deciles de ingreso**.

Distribución de los Hogares según ingreso familiar – Formosa, octubre 1997

Decil	Escala Ingresos	Hogares (%)	Ingreso total Por Decil (miles)	Porcentaje de Ingreso	Ingreso medio por decil
1	20-200	10	549	1,9	127
2	200-250	10	976	3,3	225
3	250-330	10	1281	4,3	296
4	330-400	10	1603	5,4	371
5	400-500	10	1901	6,4	439
6	500-600	10	2316	7,8	533
7	600-710	10	2796	9,4	652
8	720-980	10	3584	12,1	830
9	980-1330	10	4935	16,7	1134
10	1330-10449	10	9668	32,7	2219
Total		**100 (43288)**	**29609**	**100,0**	**684**

Fuente: INDEC – EPH. 1998

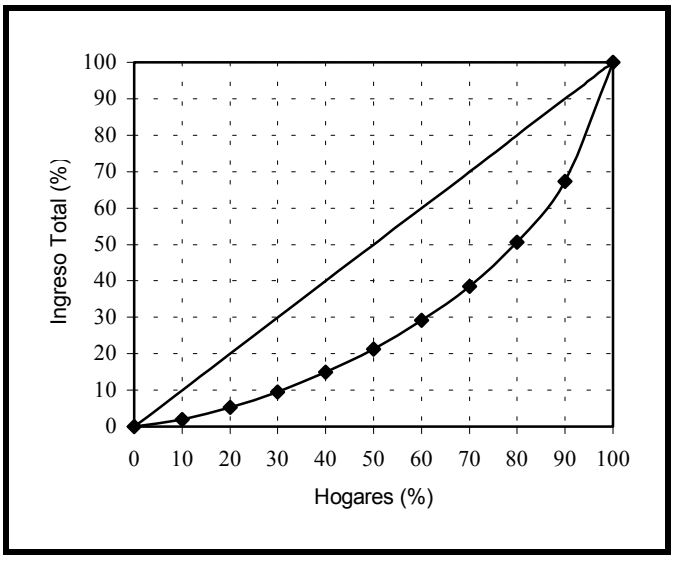
En la tabla se aprecia que, entre los hogares de la ciudad de Formosa, existe una concentración de los ingresos: el 10% de los hogares que más ganan concentran el 32,7% del total de los ingresos, mientras que el 10% de los hogares más pobres acumulan sólo el 1,9%. Esta situación produce una brecha entre "ricos" y "pobres", en la que el **ingreso promedio del último decil ($2219) es 17,5 veces mayor que el ingreso promedio del primer decil**. Esta comparación se podría extender a otros grupos, por ejemplo comparar el primer 20% de los hogares (primer quintil) que acumula sólo el 5,2% frente al último 20% que acumula el 49,4% del total de los ingresos; y así sucesivamente.

La curva de Lorenz tiene la ventaja de expresar las situaciones de equidad/inequidad de manera más general, permitiendo apreciar el comportamiento de la variable en forma inmediata.

Según hemos visto en la unidad anterior, para construir la curva de Lorenz tenemos que realizar las siguientes transformaciones: acumular los porcentajes de hogares y acumular los porcentajes de ingresos totales por decil.

Distribución de los Hogares según deciles de ingreso - Formosa, octubre 1997

Decil	Escala Ingresos	Hogares Acum. (%)	Ingresos Acum. (%)
1	20-200	10	1,9
2	200-250	20	5,2
3	250-330	30	9,5
4	330-400	40	14,9
5	400-500	50	21,3
6	500-600	60	29,1
7	600-710	70	38,5
8	720-980	80	50,6
9	980-1330	90	67,3
10	1330-10449	100	100,0



La curva así construida expresa de manera elocuente la **concentración del ingreso** que existe en los hogares de Formosa, y el hecho de haber utilizado los deciles facilita la lectura comparativa de los datos.

Actividad Nº 6

Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 6 de la Guía de Actividades correspondiente a esta unidad.

7. ¿Cómo Integrar estas Medidas de Resumen?

Hemos presentado hasta aquí una serie de valores característicos de una distribución que nos permite señalar diferentes aspectos del conjunto de datos que se analiza. Cada una de estas medidas dirige nuestra mirada hacia algún rasgo de interés de ese conjunto, las que serán más ilustrativas en tanto sean integradas en una descripción que totalice todos los aspectos destacables, generando así una "buena imagen" de esa distribución.

7.1. El resumen de los cinco números

Una forma aceptada y eficaz de integrar diferentes medidas descriptivas es la que se conoce como "**el resumen de los cinco números**", en la que se consideran:

- $\mathbf{X_{mín}}$: el mínimo
- $\mathbf{Q_1}$: el cuartil 1
- $\mathbf{M_a}$: la mediana
- $\mathbf{Q_3}$: el cuartil 3
- $\mathbf{X_{máx}}$: el máximo

Con estos valores, estamos describiendo la distribución identificando un valor de tendencia central (la mediana), dos valores entre los cuales se concentran el 50% de los datos centrales (Q_1 y *Q_3) y otros dos valores entre los cuales se dispersa el conjunto total de los datos ($X_{mín}$ y $X_{máx}$).*

Si consideramos los gastos diarios de los grupos turísticos, podemos describir mediante este criterio al conjunto de las observaciones utilizando los siguientes valores:

$\mathbf{X_{mín}}$= \$ 0 $\mathbf{Q_1}$= \$52,11 $\mathbf{M_a}$= \$101,75 $\mathbf{Q_3}$ = \$155,83 $\mathbf{X_{máx}}$= \$385

"*La mitad de los grupos turísticos no superan los $101,75 de gasto diario, aunque los gastos observados varían $0 y $385. Por otro lado, el 50% de los gastos centrales se ubican entre $52,11 y $155,83".*

Así como el resumen de los cinco números resulta un recurso apropiado para hacer una descripción de la distribución, también se pueden incorporar otros valores característicos que expresen nuevas especificidades del conjunto de datos. En este sentido, es posible agregar al análisis, otras medidas que nos permitan dar una mejor idea de la forma de la distribución. Por ejemplo, utilizando además de los cinco números vistos, los deciles 1 y 9 en un resumen que podríamos llamar "de los siete números".

$\mathbf{X_{mín}}$= \$ 0 $\mathbf{D_1}$= \$20,8 $\mathbf{Q_1}$= \$52,11 $\mathbf{M_a}$= \$101,75 $\mathbf{Q_3}$ = \$155,83 $\mathbf{D_9}$= \$231 $\mathbf{X_{máx}}$= \$385

Al comentario anterior basado en los cinco números, se podría agregar que:

"El 10% de los que menos gastan no superan los $20,8 diarios, mientras que un 10% de los grupos turísticos, gastan diariamente $231 o más".

IMPORTANTE

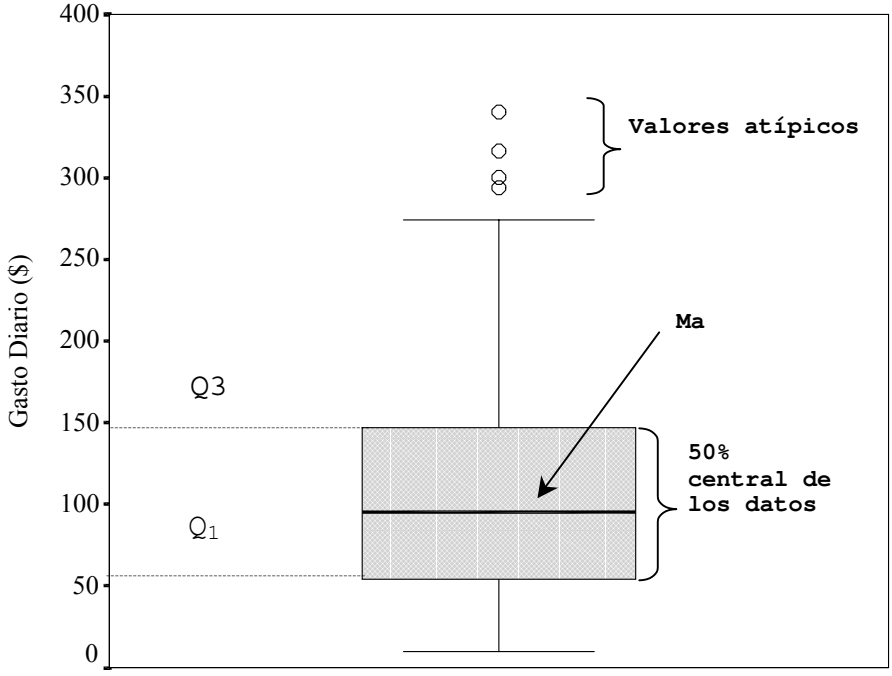
La decisión del número de valores característicos a utilizar para la descripción, e incluso qué deciles incorporar, depende de las particularidades de la distribución: número de casos, forma, número de valores diferentes que tome la variable y propósitos del análisis.

7.2. El diagrama de Caja (*Box-plot*)

El **recurso gráfico asociado al resumen de los cinco números** es lo que se conoce como Diagrama de Caja [22]. En este diagrama se utiliza un rectángulo (caja) que limitado por los cuartiles uno y tres, incluye en su interior el 50% de los datos centrales; dentro de la caja se señala la mediana con un segmento. A partir de esos límites del rectángulo, se grafican líneas -llamadas "bigotes"- con una longitud igual a 1,5 veces la distancia entre el cuartil 1 y el 3 [23]. Posteriormente –fuera de los "bigotes"- el gráfico identifica aquellos valores atípicos (*outliers*), que están a más de 1,5 veces la distancia Intercuartil (1,5 . R Q) de los extremos de la caja.

A continuación presentamos el diagrama de Caja construido a partir de los datos individuales de los gastos realizados diariamente por los 72 grupos turísticos.

Diagrama de Caja: Distribución de los gastos diarios.
Pto Iguazú, Feb. '94



En este gráfico podemos ver que los gastos diarios de los turistas tienen un comportamiento bastante simétrico en el 50% de los datos centrales (la mediana se ubica en el centro de la caja, a igual distancia de los cuartiles uno y tres). El conjunto total de los datos muestra una asimetría a la derecha, (el bigote superior es más largo que el inferior e incluso se aprecia la presencia de cuatro grupos turísticos con gastos atípicos). Por otro lado el "bigote" inferior está indicando una mayor concentración de los gastos menores, no hay valores atípicos pequeños e incluso no se identifica ningún grupo que no haya realizado gastos (el "bigote" no alcanza al valor $0).

Este tipo de recurso gráfico resulta muy ilustrativo y en consecuencia recomendable cuando queremos comparar dos o más distribuciones [24].

Vemos entonces que el diagrama de caja permite visualizar una serie de aspectos interesantes de la forma del conjunto de los datos:

- *Presencia de* ***valores atípicos***

[22] También denominado "Diagrama de Caja con bigotes" o en inglés "*Box-Plot*".

[23] En la unidad siguiente, se podrá ver que esta distancia entre el cuartil 1 y el 3 es una medida de variabilidad que se conoce como *Rango intercuartil* (RQ).

[24] El uso del *box-plot* para la comparación de conjuntos de datos, será tratado posteriormente en la Unidad 5.

- ***Simetría del conjunto central*** *de los datos (equidistancia o no de la mediana a los cuartiles).*
- ***Simetría del conjunto total*** *de datos (forma de la caja y longitud de los bigotes).*
- ***Dispersión en cada una de las zonas*** *en las que queda dividida la distribución (la longitud de cada parte, expresa la mayor o menor proximidad de los datos entre sí).*
- *El* ***rango*** *de la distribución (distancia entre el valor máximo y mínimo).*

Estas características del diagrama hacen que el mismo resulte **útil** (junto con el de *tallo-hoja*) en la **etapa inicial exploratoria** de los datos, previo a la construcción de una distribución de frecuencias y cálculo de las medidas resumen, ya que -como hemos visto- **la forma de la distribución condiciona el posterior tratamiento y resumen de los datos**.

Actividad Nº 7

*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 7 de la Guía de Actividades correspondiente a esta unidad***.**

8. ¿Qué Hemos Visto?

En esta unidad hemos avanzado un paso más en el camino del tratamiento y análisis estadístico elemental de los datos.

Efectuados los primeros resúmenes numéricos y gráficos, para una primera lectura del fenómeno que representan los datos (unidad 2), el análisis a menudo requiere de instrumentos que permitan un **mayor resumen de la información**.

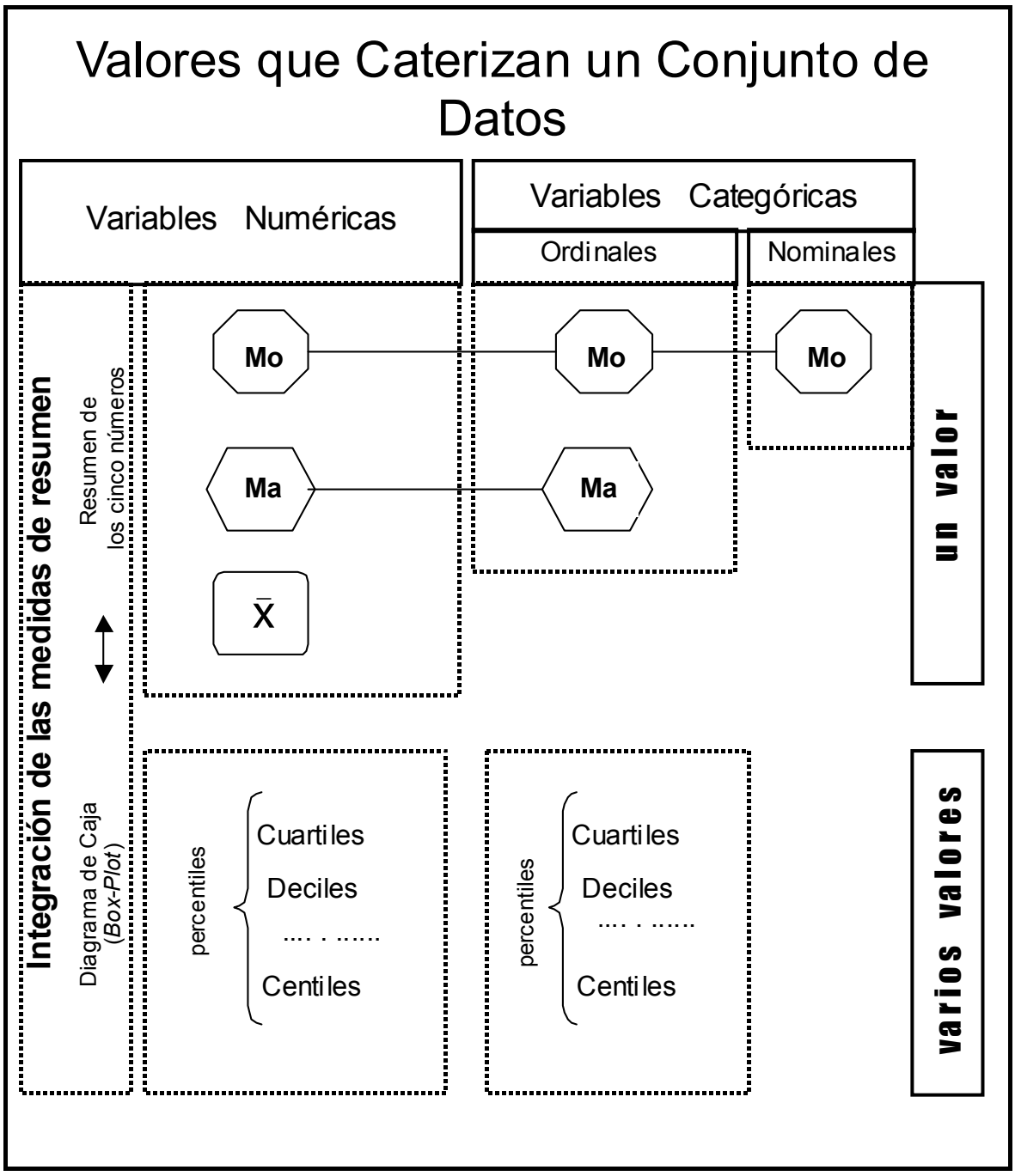
Las medidas de **tendencia central** tienen este propósito, y su aplicación en un problema particular **dependerá básicamente de las necesidades de información que motivan el análisis**, del **tipo de datos** con los que se trabaja y de las **propiedades del conjunto** como un todo.

El buen dominio del concepto, propiedades y limitaciones de cada una de ellas es el requisito para utilizarlas correctamente.

Además, hemos presentado las **diferentes medidas de posición** que permiten complementar la comprensión de un conjunto de datos, informando -con distintos niveles de detalle- sobre su estructura.

En todos los casos, el énfasis está puesto en facilitar la comprensión conceptual de cada herramienta, para luego pasar al plano de la formalización matemática elemental y del cálculo aplicado a ejemplos de fácil comprensión.

En relación con esto último, reiteramos la recomendación a quienes puedan hacerlo, de utilizar la informática como auxiliar del trabajo estadístico.


Valores que Caterizan un Conjunto de Datos
Variables Numéricas
Variables Categóricas
Ordinales
Nominales
Integración de las medidas de resumen
Resumen de los cinco números
Diagrama de Caja (Box-Plot)
Mo
Mo
Mo
Ma
Ma
X̄
un valor
percentiles
Cuartiles
Deciles
....
Centiles
percentiles
Cuartiles
Deciles
....
Centiles
varios valores

Bibliografía

BARBANCHO, A. (1978): *Estadística Elemental Moderna*. Ed. Ariel, Barcelona, España. pág. 117-123, 127-132, 134-138.

BLALOCK, H. M.(1978): *Estadística Social,* FCE, México. pág. 67-72, 81-83.

UNIVERSIDAD NACIONAL DE CÓRDOBA (1993): *Estadística aplicada a la Investigación. Curso a distancia*. Fac. de Cs. Económicas, Córdoba, Módulo III pág. 1-42.

Conceptos Centrales

- Media aritmética: concepto y propiedades.
- Mediana: concepto y propiedades.
- Modo: concepto y propiedades.
- Cuartiles, deciles, centiles: concepto y aplicación.

Habilidades

- Reconocer la utilidad, alcances y limitaciones de cada una de las medidas resumen presentadas.
- Identificar para una situación de trabajo, las medidas de Tendencia Central y Posición que podrían utilizarse para una buena descripción de los datos.
- Conocer los fundamentos que guían los procedimientos para la obtención de estas medidas.
- Interpretar en términos de un problema, las medidas y gráficos asociados a una distribución (*Box-plot* y Curva de Lorenz).
- Saber comunicar en un informe las características de un conjunto de datos, integrando los distintos recursos estadísticos aprendidos hasta el momento.