UNIDAD 4: ANÁLISIS DE LA VARIACIÓN Y ASIMETRÍA

1. ¿Por qué Evaluar la Variabilidad y la Asimetría?

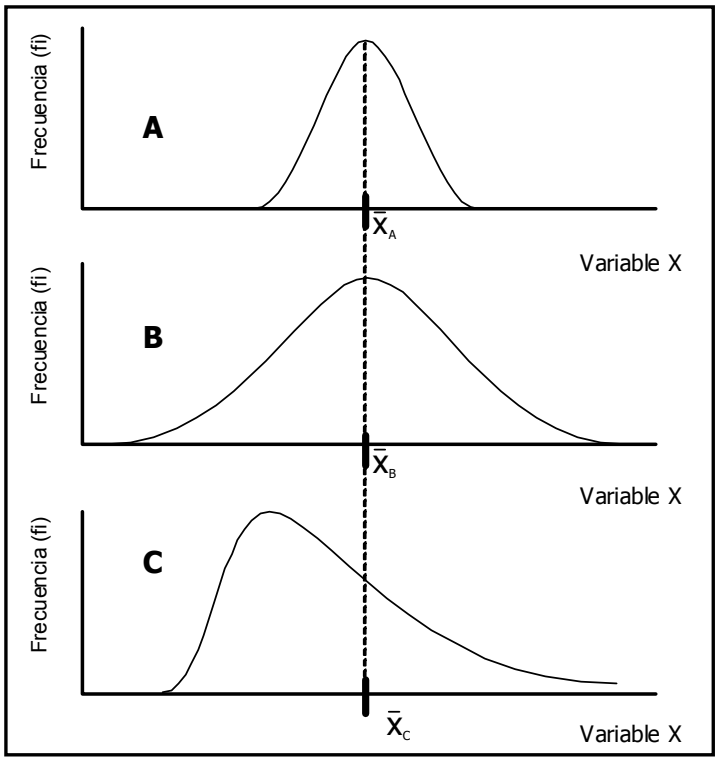
No se investiga lo obvio, aquello que encuentra una respuesta simple y evidente. Las preguntas que nos formulamos generalmente aluden a situaciones complejas, comprenden fenómenos en los que las características de interés presentan valores diversos, no son uniformes.

Dicho en términos estadísticos, los datos que obtenemos en relación con alguna pregunta de investigación, **varían a través del conjunto de unidades observadas**, y "controlar" esa variabilidad es el fin último en la tarea de describir los datos y producir información.

Hasta aquí todas las medidas o herramientas presentadas intentaban, de diferentes maneras, resumir los datos para lograr una mejor **descripción de esa diversidad**. Así, las **distribuciones de frecuencias** (en su forma numérica o gráfica) nos permiten ***presentar y describir los diferentes valores observados***. En tanto que las ***medidas resumen*** desarrolladas en la unidad anterior, nos facilitan la ***descripción de los individuos a través de un conjunto de valores característicos*** que intentan dar cuenta de la variabilidad.

Asimismo, debemos destacar que **la representatividad de las medidas de tendencia central se vincula estrechamente con la dispersión de los datos** y (concretamente en el caso de la media) con la ***simetría*** de la distribución[1]**.** Consideremos los siguientes gráficos donde se representan tres distribuciones de frecuencias (polígonos A, B y C) que registran un mismo valor para la media.

Distribuciones con igual media aritmética y diferente variabilidad y/o simetría



Evaluando los gráficos, es posible concluir que la media aritmética resulta mucho más representativa del conjunto de datos en la distribución **A** (simétrica y con menor variabilidad) que en las situaciones **B** (simétrica pero con valores más dispersos en torno a la media) y **C** (también más dispersa y asimétrica).

[1] Esto pone de manifiesto que tanto la variabilidad como la asimetría de la distribución son aspectos a considerar a la hora de evaluar estas medidas. Recordar que: cuando se observa la presencia de valores atípicos, el ***promedio aritmético debe ser analizado con cuidado,*** porque puede resultar fuertemente desplazado de la tendencia central e inducir a interpretaciones erróneas acerca del conjunto de datos que resume (Ver Unidad 3).

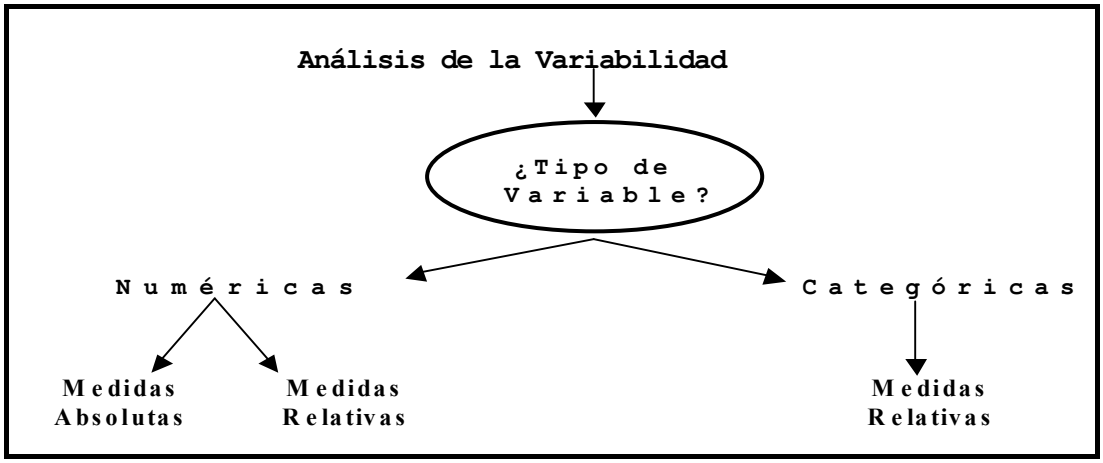
A estas características que hacen a *la forma* (variabilidad y simetría[2]) de la distribución, le podemos asociar *medidas que resuman en números la "cantidad de variación" y el "grado de asimetría"*, valores que nos permitirán comparar distintos conjuntos de individuos.

En esta unidad abordaremos -en primera instancia- cómo medir la variabilidad, para posteriormente presentar aquellas medidas del grado de asimetría de una distribución.

2. ¿Cómo Medir la Variabilidad?

¿Qué significa medir la variabilidad? Obtener un único número que exprese qué tan ***dispersos o diferentes*** son entre sí el conjunto de valores observados o -lo que es lo mismo- que indique cuán homogéneos son los individuos en términos de la característica en cuestión.

Si bien el concepto de variabilidad es único, las medidas son distintas según se trate de variables numéricas o categóricas. Además, para las variables numéricas podemos identificar medidas absolutas y relativas.



2.1. Para variables numéricas

Por tratarse de variables *medidas en una escala de intervalo*, la dispersión de los valores observados se puede expresar directamente por la diferencia aritmética entre esos valores. En consecuencia, cuanto mayor sea la diferencia entre dos valores, podemos aseverar que mayor será la variación que existe entre esos dos datos.

Veamos en un sencillo ejemplo, las ideas anteriores: tenemos seis individuos para los cuales se han registrado sus notas en Historia y Matemática.

Simboliza al segundo individuo

Individuo	i_1	i_2	i_3	i_4	i_5	i_6	**Media**
Nota Historia	7	8	7	6	7	7	**7**
Nota Matemática	4	8	5	9	10	6	**7**

Es la nota de Matemática del cuarto individuo

Se puede observar que los promedios de las notas en estas materias son coincidentes. Sin embargo, la variabilidad en las notas de Historia es claramente menor que en las de Matemática; así la mayor variación que se registra entre las notas de Historia es de 2 puntos (entre $\mathbf{i_2}$ y $\mathbf{i_4}$, que son los individuos más diferentes entre sí), mientras que en Matemática, la mayor diferencia es de 6 puntos (entre $\mathbf{i_5}$ y $\mathbf{i_1}$). Estamos en condiciones de afirmar para este pequeño conjunto de observaciones que, a pesar de que la medida resumen es la misma, los conjuntos son diferentes: las notas de Matemática

[2] Aunque no lo desarrollaremos en este curso, otro aspecto a considerar en el análisis de la forma es lo que se conoce como *curtosis*.

son más heterogéneas (están más dispersas) que las de Historia. El promedio en Historia ***"representa mucho mejor"*** al rendimiento de los estudiantes en esa asignatura, que la nota promedio de Matemática al correspondiente conjunto de datos.

IMPORTANTE

Las medidas de tendencia central ocultan la variabilidad del conjunto de datos. Por ello, cuantificar la variabilidad constituye un complemento imprescindible en la descripción de una distribución.

Conocer (medir) la variación de los datos permite:

- ➔ describir esta característica inherente a todo conjunto de observaciones,
- ➔ evaluar la *"calidad"* de las medidas de tendencia central, y
- ➔ comparar mejor diferentes grupos de datos mediante sus promedios.

En general, las situaciones no serán tan evidentes, ni el número de datos tan pequeños como en el ejemplo anterior; lo que obliga a construir medidas que nos permitan resumir y evaluar esa variabilidad.

2.1.1. Las medidas absolutas

Para la construcción de medidas absolutas de variación se pueden adoptar dos perspectivas:

➔ **Considerar el campo de variación de las variables**: las medidas obtenidas expresan la extensión o amplitud de variación de los datos que se están considerando. Se identifican en este grupo: el *Rango* y el *Rango Intercuartil*.

➔ **Considerar las variaciones de los datos individuales:** estas medidas resumen en un valor la totalidad de las variaciones de los datos individuales. Entre estas medidas se destacan: la *Desviación Media*, la *Desviación Mediana*, la *Variancia* y el *Desvío Estándar*.

Considerando el campo de variación de las variables, tenemos:

A) El Rango, Amplitud o Recorrido: indica la extensión en la que varían la totalidad de los datos; es la mayor diferencia que se puede registrar entre dos valores de la variable.

Esta medida se calcula como la diferencia entre el máximo valor y el mínimo valor observado de la variable.

$$R = x_{máx} - x_{mín}$$

En el ejemplo de las notas el rango para la variable "nota de Matemática" es de 6 (R = 10 - 4), lo que indica que la totalidad de las notas observadas se registran en un campo o extensión de variación de 6 puntos. En el caso de las "notas de Historia" esta amplitud de variación es de 2 puntos.

Cuando los datos están agrupados en intervalos de clase, dado que no conocemos exactamente el máximo y el mínimo, el rango se obtiene [3] haciendo la diferencia entre el límite superior de la última clase y el límite inferior de la primera:

$$R = L_{sk} - L_{í1} \quad \text{(donde k es el número de clases)}$$

Comentarios*:*

Es una medida de muy fácil cálculo, que permite una aproximación rápida a la variabilidad de los datos.

Al tomar sólo los valores máximo y mínimo, si se observan **valores muy atípicos**, puede brindar una **idea distorsionada** sobre la variabilidad como característica del conjunto.

Dos distribuciones con el **mismo rango** pueden tener dispersión "interna" de los **datos muy diferentes** (el conjunto de los valores pueden estar más o menos concentrados).

[3] Estrictamente se trata de una estimación ya que desconocemos los verdaderos valores máximos y mínimos.

B) Rango intercuartil: indica la extensión en la que varían el 50% de los datos centrales de la distribución.

Se calcula como la diferencia entre el tercer y el primer cuartil.

$$RQ = Q_3 - Q_1$$

Comentarios:

Muchas veces es preferible medir la variabilidad del 50% de los datos centrales, descartando el 25% de los valores más bajos y el 25% de los más altos, para evitar así la distorsión que puede provocar la presencia de valores atípicos.

Simultáneamente, estamos prescindiendo en este caso de la mitad de las observaciones.

Para describir la distribución de las edades de los alumnos del curso de Estadística podemos utilizar algunas de las medidas de resumen presentadas en la unidad anterior.

Mediana	21 años
Mínimo	17 años
Máximo	47 años
Cuartil 1	19 años
Cuartil 3	27 años

A estas medidas las podemos complementar con medidas de variación. Así tenemos:

Rango: R = 47 - 17 = 30 años

Rango intercuartil: RQ= 27 - 19 = 8 años

A partir de este conjunto de medidas se puede decir que: *la mitad de los alumnos de Estadística tienen 21 años o menos, y los más jóvenes tienen 17 años. Las edades de los estudiantes varían en una amplitud de 30 años, lo que implica una diferencia de 30 años entre el/(los) alumno/s más jóven/es y el/(los) de más edad. El 50% de los estudiantes con las edades centrales difieren a lo sumo en 8 años.*

Recordar que el Diagrama de Caja (*Box-Plot*) es un recurso gráfico apropiado para el análisis de la distribución en general y de la variabilidad y asimetría en particular. (Ver unidad 3).

Actividad Nº 1

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

Considerando las variaciones de los datos individuales tenemos:

Una alternativa que facilita construir estas medidas es tomar los ***desvíos de cada uno de los valores individuales con respecto a un punto elegido como referencia***. Generalmente este valor de referencia es una medida de tendencia central.

C) Desviación media: esta medida se construye tomando ***todos*** **los** ***desvíos individuales*** con respecto a la media aritmética.

Como hemos definido, *un* ***desvío individual*** *es la diferencia entre un valor de la variable y la media aritmética*: $d_i = (x_i - \bar{x})$. Es decir que tendremos tantos desvíos individuales como individuos hayamos observado.

En el ejemplo de las notas de Matemática tendríamos los seis desvíos siguientes:

Individuo	i_1	i_2	i_3	i_4	i_5	i_6	**Media**
Nota Matemática	4	8	5	9	10	6	**7**
Desvíos individuales a la $\bar{x}$	-3	1	-2	2	3	-1	

Se puede ver que, mientras el individuo 1 está 3 puntos por debajo de la media, el individuo 5 está en esa misma cantidad por encima de la media.

Para resumir en un único número la variabilidad de las seis observaciones, podemos recurrir al promedio pero, como ya hemos señalado en la unidad anterior, *la suma de los desvíos a la media es cero* [4]. Para resolver este problema vamos a sumar los desvíos absolutos, es decir el valor de los desvíos prescindiendo de su signo.

En términos del problema tenemos que la *Desviación Media* se obtiene como:

$$DM = \frac{3+1+2+2+3+1}{6} = \frac{12}{6} = 2 \text{ puntos}$$

Se interpreta que, *en promedio, las notas de matemática se desvían de la media en 2 puntos*.

Desviación Media (DM):

Es el promedio de los desvíos individuales (en valor absoluto) con respecto a la media aritmética.

$$DM = \frac{\sum |x_i - \bar{x}|}{n} = \frac{\sum |d_i|}{n}$$

Las barras simbolizan "valor absoluto"

Comentario:

Cuando estamos en presencia de distribuciones en las que se observan **valores atípicos** (marcadamente asimétricas) la media como medida resumen *no es recomendable*, y en consecuencia *tampoco lo es la desviación media* como medida de variabilidad.

Para el caso de las edades de los alumnos del curso de Estadística, la *Desviación Media* calculada a partir de los valores individuales, es: DM = 5,14 años (Ud. podría controlar este resultado, calculando la DM a partir de los datos que figuran en la Unidad 2).

Para datos organizados en una distribución de frecuencias:

- Si se trata de un **arreglo** de frecuencias y se va a obtener la desviación media en forma manual, la expresión de cálculo es:

$$DM = \frac{\sum |x_i - \bar{x}| . f_i}{n} = \frac{\sum |d_i| . f_i}{n}$$ donde f_i es la frecuencia del valor x_i

- Cuando los datos están *agrupados en* ***intervalos de clase***, *y no se dispone de los valores individuales*, se podrá estimar la Desviación Media, considerando que el x_i de la fórmula se corresponde con el punto medio de la clase.

En el caso de la edad de los estudiantes, si desconociéramos los valores individuales de esta variable y contáramos únicamente con los datos organizados en una distribución de frecuencias en intervalos de clase, podríamos estimar la Desviación Media realizando las operaciones que se indican en la Tabla.

Estudiantes del curso de Estadística según edad- FHyCS-Año 2001

Punto Medio de clase — **Desvíos individuales**

Edad	nº de estud. (f_i)	PM	d_i = (PM– 23,6)	$\|d_i\| . f_i$
17-20	65	18,5	-5,1	331,5
21-24	25	22,5	-1,1	27,5
25-28	17	26,5	2,9	49,3
29-32	14	30,5	6,9	96,6
33-36	7	34,5	10,9	76,3
37-40	5	38,5	14,9	74,5
41-44	2	42,5	18,9	37,8
45-48	1	46,5	22,9	22,9
Total	**136**			**716,4**

Fuente: elab. propia en base a datos del "Estudio de los Alumnos de Estadística"

[4] Recordar que por una propiedad de la media la suma de los desvíos individuales a la media siempre es cero. $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

Dividiendo la suma de los desvíos en valores absolutos (716,4) por el número de casos (136), tenemos una Desviación Media estimada en 5,27 años.

"Las edades de los alumnos de estadística se dispersan –en promedio- con respecto a la media en 5,27 años".

D) Desviación mediana: si evaluamos que la media no es una buena medida resumen de los datos y optamos por la mediana como medida de tendencia central, sería apropiado utilizar una medida de dispersión relacionada a la mediana. Así entonces, de manera análoga a la desviación media, tenemos que:

Desviación Mediana (DMa):

Es el promedio de los desvíos individuales (en valor absoluto) con respecto a la mediana.

$$DMa = \frac{\sum |x_i - Ma|}{n}$$

Comentarios:

- Para datos organizados en distribuciones de frecuencias, valen los mismos comentarios que para el cálculo de la Desviación Media.

$$DMa = \frac{\sum |x_i - Ma| . f_i}{n}$$

donde:

f_i es la frecuencia del valor x_i

x_i son los valores observados de la variable en el caso de un arreglo de frecuencias, o el punto medio de la clase en el caso de una distribución en intervalos de clase.

Calculamos la Desviación Mediana para las Notas de Matemática:

Individuo	i_1	i_2	i_3	i_4	i_5	i_6	Ma
Nota Matemática	4	8	5	9	10	6	7
Desvíos a la Ma	-3	1	-2	2	3	-1	

Promedio de los valores centrales 6 y 8

$$DMa = \frac{\sum |x_i - Ma|}{n} = \frac{3+1+2+2+3+1}{6} = 2 \text{ puntos}$$

En consecuencia, *las notas de Matemática se desvían, en un promedio de 2 puntos, de la mediana*.

Estudiantes del curso de Estadística según edad- FHyCS-Año 2001

La edad de los estudiantes es una distribución marcadamente asimétrica a la cha y la mediana (Ma = 21,5) será la mejor medida resumen de los datos. Así, lo más apropiado es utilizar la desviación mediana, que se obtiene mediante las operaciones que se presentan en la Tabla:

Edad	n° de estud. (f_i)	PM	d_i = (PM – 21,5)	$\|d_i\| . f_i$
17-20	65	18,5	-3	195
21-24	25	22,5	1	25
25-28	17	26,5	5	85
29-32	14	30,5	9	126
33-36	7	34,5	13	91
37-40	5	38,5	17	85
41-44	2	42,5	21	42
45-48	1	46,5	25	25
Total	136			674

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística"

Desvíos individuales a la mediana

Suma del producto de los **desvíos absolutos** individuales a la mediana por la frecuencia

Luego: $DMa=\frac{674,0}{136}=4,96$ años

"Esta medida indica que en promedio las edades de los estudiantes se desvían de la mediana en 4,96 años".

E) Variancia y Desviación estándar: en el cálculo de la desviación media se tomaron los valores absolutos de los desvíos evitando así que la suma nos dé cero. Otro criterio para solucionar este mismo problema sería elevar esos desvíos al cuadrado, obteniendo de esta manera una nueva medida de variabilidad que se conoce como Variancia.

Esta medida se simboliza utilizando la letra griega "sigma" elevada al cuadrado (σ^2).

El cálculo de la variancia para las notas de Matemática es:

Individuo	i_1	i_2	i_3	i_4	i_5	i_6	**Media**
Nota Matemática	4	8	5	9	10	6	**7**
Desvíos individuales a la $\bar{x}$	-3	1	-2	2	3	-1	

¿. . . ?

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{(-3)^2 + (1)^2 + (-2)^2 + (2)^2 + (3)^2 + (-1)^2}{6} = \frac{28}{6} = 4,7 \text{ (puntos)}^2$$

Variancia **(σ^2)**:

Es el promedio de los cuadrados de los desvíos a la media aritmética.

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Comentarios:

- La *variancia* y el *desvío estándar* son, fundamentalmente por razones de orden teórico, las medidas *más utilizadas* para cuantificar la variabilidad de un conjunto de datos.
- Dado que los desvíos a la media están elevados al cuadrado, la variancia ***se expresa en una unidad de medida que es el cuadrado de la unidad de medida de la variable original***. Esto dificulta la interpretación del resultado en términos del problema.

La unidad de medida en la que queda expresada la variancia no es interpretable en términos de la variable que se analiza. Hasta aquí sólo la podemos considerar como una cuantificación de la variabilidad existente en los datos.

Para resolver este problema, se calcula la *raíz cuadrada de la variancia*, que resulta en una nueva medida llamada ***Desvío Estándar*** (σ), la que queda expresada en la unidad original.

$$\sigma = \sqrt{\sigma^2}$$

En el ejemplo de las notas de Matemática el desvío estándar será:

$$\sigma = \sqrt{4,7} = 2,2 \text{ puntos}$$

"*Las notas de matemática de los alumnos se dispersan en promedio en torno a la media en 2,2 puntos*".

Si no contáramos con los datos originales, el cálculo de la variancia y el desvío estándar para las edades de los estudiantes de estadística, a partir de la tabla, sería:

Estudiantes del curso de Estadística según edad- FHyCS-Año 2001

Desvíos individuales a la media

Desvíos al cuadrado

Edad	nº de estud. (f_i)	PM	d_i = (PM– 24,1)	d_i^2	$d_i^2 . f_i$
17-20	65	18,5	-5,1	26,0	1690,0
21-24	25	22,5	-1,1	1,2	30,0
25-28	17	26,5	2,9	8,4	142,8
29-32	14	30,5	6,9	47,6	666,4
33-36	7	34,5	10,9	118,8	831,6
37-40	5	38,5	14,9	222,0	1110,0
41-44	2	42,5	18,9	357,2	714,4
45-48	1	46,5	22,9	524,4	524,4
Total	**136**				**5709,6**

Fuente: elaboración propia basada en datos del "Estudio de los Alumnos de Estadística"

La variancia es: $\sigma^2 = \frac{\sum d_i^2 . f_i}{n} = \frac{5709,6}{136} = 42,0$

El desvío estándar es: $\sigma = \sqrt{42} = 6,48$ años.

Entonces, *los estudiantes del curso tienen una media de 24,1 años y sus edades -en promedio- se dispersan con respecto a ese valor 6,48 años*.

Para datos agrupados en distribuciones de frecuencias:

La expresión de cálculo es:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 . f_i}{n}$$

donde: $\mathbf{f_i}$ es la frecuencia del valor x_i

$\mathbf{x_i}$ son los valores observados de la variable en el caso de un arreglo de frecuencias, o el punto medio de la clase en el caso de una distribución en intervalos de clase.

Actividad Nº 2

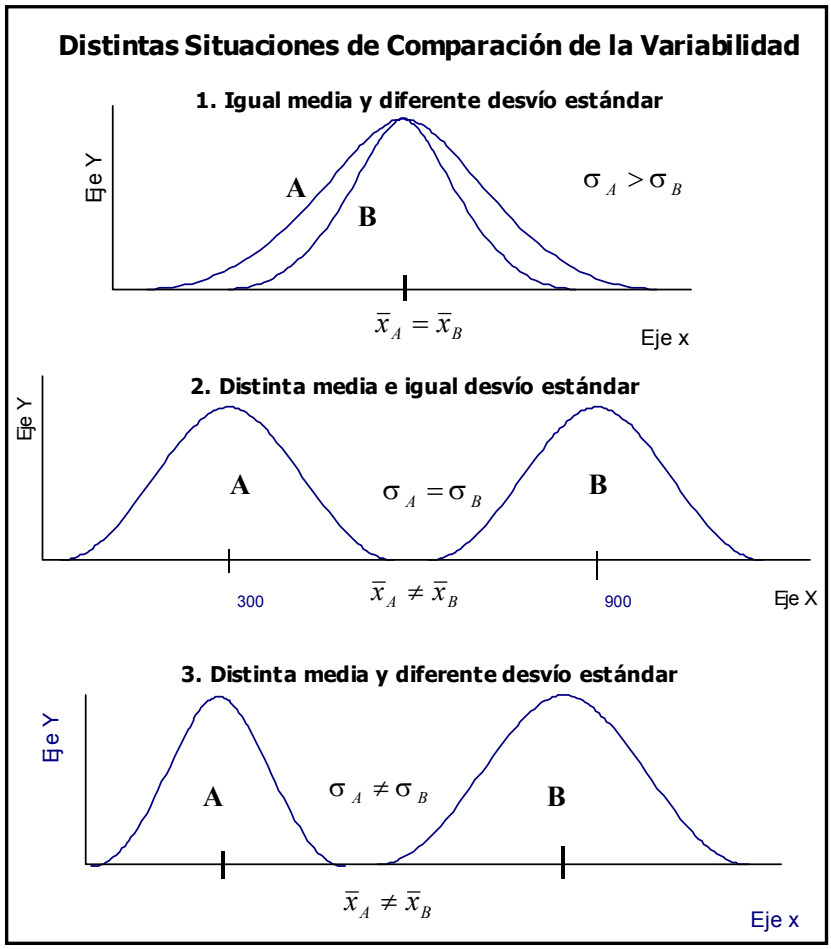
*Antes de continuar con la lectura, es necesario realizar aquí la Actividad Nº 2 de la Guía de Actividades correspondiente a esta unidad***.**

2.1.2. Las medidas relativas

Con frecuencia nos vemos en situaciones de tener que ***comparar la variabilidad de diferentes conjuntos de datos***. Así por ejemplo, comparar los ingresos de grupos pertenecientes a distintos estratos sociales, las edades de grupos en diferentes etapas de la vida, las temperaturas en distintas regiones del planeta, etc.

Existen **diferentes situaciones** que se pueden presentar **al comparar distribuciones**. En el esquema siguiente se presentan, en términos generales, esas situaciones de comparación.

El **primer Gráfico** está expresando una situación en la cual debemos comparar la variabilidad de dos grupos que -medidos en la misma variable- tienen **medias iguales** y **dispersiones diferentes**. Es fácil de concluir que en la distribución B los individuos son más homogéneos que en la otra.



La dificultad de comparar no se presenta tan clara en las otras dos situaciones (2 y 3).

Cuando las variables están medidas en la misma escala (situación 2), no es difícil de ver que:

- una variación de 2 años entre escolares, no implica la misma heterogeneidad de los individuos (en cuanto a: intereses, preferencias y habilidades) que esa misma variación entre universitarios, o
- una dispersión de $50 pesos en el ingreso mensual de gerentes de empresa, no los diferencia (en cuanto al nivel de vida o consumo), de la misma manera que esa misma variación lo hace entre sus obreros, etc.
- Es aún más evidente la dificultad de comparar la hogeneineidad de los individuos cuando las distribuciones tienen valores distintos de promedio y dispersión absoluta (situación 3). Por ejemplo esto ocurriría si queremos comparar:
    - la variación en el consumo de energía eléctrica de los hogares y de las industrias. Si conociéramos que el desvío estándar en el consumo de los hogares es de 100 Kw y entre las industrias es de 1500 Kw; no tenemos información suficiente para concluir sobre la mayor o menor homogeneidad en alguno de las poblaciones, dado que -como podemos suponer- sus promedios son sustancialmente diferentes.

En consecuencia, **para valorar la dispersión de un grupo y poder compararlo con otro**, se hace ***necesario evaluar la dispersión en términos relativos a las magnitudes de esas variables en cada uno de los grupos***. Esto significa que, comparar la cantidad de dispersión de dos grupos, exige construir ***medidas relativas de variabilidad***.

Esta necesidad de ***relativizar la variabilidad***, se evidencia también cuando se busca comparar la homogeneidad de dos conjuntos de observaciones en términos de dos ***variables expresadas en unidades de medida distintas***. Por ejemplo, queremos ver si nuestros estudiantes se parecen más entre sí (son más homogéneos) en cuanto al tiempo que miran televisión (en horas), que en relación a su edad (en años); los turistas que visitan Puerto Iguazú se parecen más entre sí en términos de sus años de estudio que de sus gastos, etc. Así, los interrogantes nos conducirían a comparar la dispersión de la edad de los alumnos con la dispersión en el tiempo que miran TV; y la variabilidad de gastos de los turistas, con la variabilidad en los años de estudio. Ambas situaciones son incomparables en términos de variabilidad absoluta.

F) Coeficiente de variación

Es la medida relativa de dispersión más utilizada dado que se construye a partir de la desviación estándar que, como hemos dicho, es la medida de dispersión más difundida.

Coeficiente de Variación (CV):

Definido como:

$$CV = \frac{\sigma}{\bar{x}}.100$$

indica la cantidad de variación expresada como un porcentaje de la media aritmética.

Comentarios:

- Si las medias aritméticas de dos conjuntos son iguales (o aproximadamente) las medidas absolutas serán suficientes para la comparación.

	Edad	**Hs. TV**
N	136	139
$\bar{X}$	23,4 años	2,0 hs.
σ	6,4 años	1,5 hs.
CV	27,3 %	75,8 %

En el ejemplo de los estudiantes, podemos ver que las *edades se dispersan en promedio un 27,3% del valor de la media aritmética*, mientras que el *tiempo que miran TV tiene una dispersión del 75,8% del promedio general*. En conclusión, "*el grupo es mucho más homogéneo en términos de sus edades que en relación con sus hábitos como televidentes".*

Existen **otras medidas relativas** de variación que se construyen de manera análoga al coeficiente de variación, según sea la medida absoluta de dispersión que se considere. Así tenemos:

G) Coeficiente de Desviación Media

$$CDM = \frac{DM}{\bar{x}}.100$$

H) Coeficiente de Desviación Mediana

$$CDMa = \frac{DMa}{Ma}.100$$

donde: *DM* es la desviación media y *DMa* es la desviación mediana.

IMPORTANTE

En la práctica *no se construyen sucesivamente* todas las medidas que hemos presentado sino que, a partir de la medida de resumen seleccionada como más representativa de la ***tendencia central, se seleccionará una medida de dispersión que la complemente***, y consecuentemente se construirá la medida relativa correspondiente a esa medida absoluta.

Una vez más: la **utilización de determinadas medidas** es el resultado de una **decisión del investigador** y surge de considerar las características de ese particular conjunto de datos que se está analizando.

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 **.**

2.2. Dispersión para variables categóricas

Como es de suponer, la construcción de una ***medida de dispersión*** para variables categóricas (nominales u ordinales) no se basa en el desvío de los datos individuales a una medida de tendencia central; su lógica es totalmente diferente. En estos casos, ***¿cómo entenderíamos y valoraríamos diferentes situaciones de dispersión?***

Supongamos que se observan seis "individuos" en una variable con dos categorías: Cat1 y Cat2 de una escala nominal u ordinal. Tendríamos así situaciones de:

- **Dispersión Nula (máxima concentración**): cuando todas las observaciones corresponden a ***una sola de las categorías*** posibles. Es decir alguna de las siguientes dos situaciones.

Situación A

Variable	nº individuos
Cat1	6
Cat2	0
Total	6

Todos los individuos presentan la característica Cat1

Situación B

Variable	nº individuos
Cat1	0
Cat2	6
Total	6

Todos los individuos presentan la característica Cat2

- **Máxima Dispersión (Mínima Concentración**) las observaciones se distribuyen entre las diferentes categorías de manera tal que, en todas, haya la misma cantidad de casos.

Variable	nº individuos
Cat1	3
Cat2	3
Total	6

- **Dispersión intermedia**: Cuando las observaciones se distribuyen entre las categorías de modo desigual pero sin llegar al extremo de concentrarse todas en una sola de ellas. Por ejemplo; situaciones como las siguientes:

Variable	nº individuos
Cat1	4
Cat2	2
Total	6

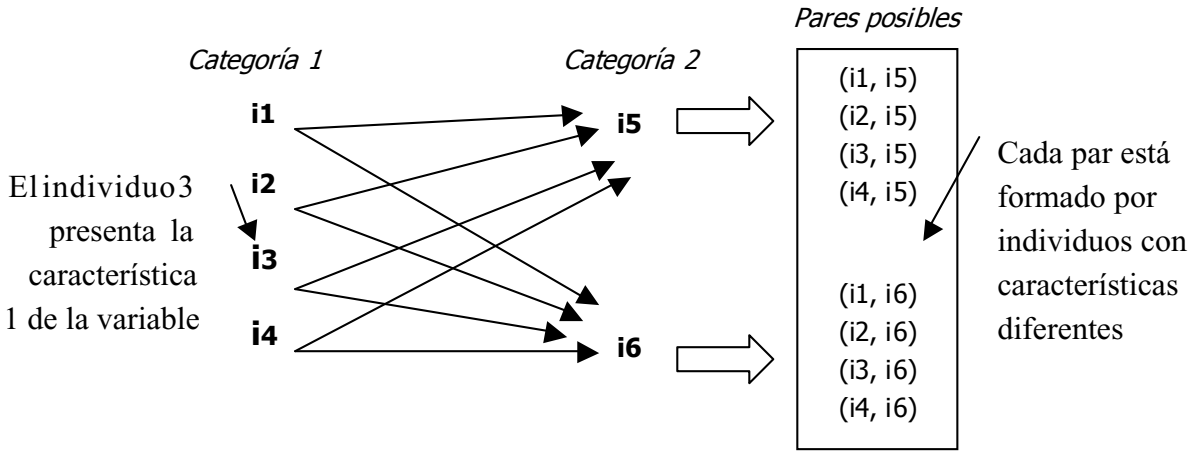
Algunas de las categorías tiene más casos que las otras

Variable	nº individuos
Cat1	1
Cat2	5
Total	6

A partir del concepto de dispersión para datos categóricos, podemos ver la ***lógica que sirve de base para la construcción del Índice de Dispersión***.

El índice de dispersión para una variable de *dos categorías* se obtiene a partir del número de pares de individuos [5] que se pueden construir combinando los elementos de una categoría con todos los de otra. Hay que tener en cuenta que, en este caso, **cada par es una combinación de individuos diferentes** en términos de la variable que se está analizando. Por ejemplo, si se tratara de la variable sexo, cada par estaría integrado por un hombre y una mujer. Así, para una variable cuya distribución presenta cuatro individuos en una categoría y dos en la otra, los pares que se pueden formar serían:

[5] Si la variable tiene tres categorías serán ternas, si tiene cuatro serán grupos de cuatro individuos y así siguiendo.



En la tabla siguiente resumimos, para el ejemplo de seis observaciones en una variable de dos categorías, el número de ***pares posibles de elementos con atributos diferentes que se pueden construir para cada nivel de dispersión***.

Nivel Dispersión	nº individuos en Cat1	nº individuos en Cat2	Nº pares posibles
Nula	6	0	0
Intermedia 1	5	1	5
Intermedia 2	4	2	8
Máxima	3	3	9

En la tabla anterior se puede ver que, *a medida que **crece el nivel de dispersión** de la variable, **aumenta el número de pares posibles** a construir.*

Se observa que la situación de máxima dispersión se corresponde con el mayor número de pares posibles y que la dispersión nula no permite construir ningún par. En consecuencia, el número de pares de diferentes elementos podría constituir una medida absoluta de la heterogeneidad de los individuos en términos de la variable en estudio.

Es posible entonces usar esta relación para construir una *medida relativa* de *dispersión*, de tal manera que sea útil para comparar distintas distribuciones.

Indice de Dispersión (ID)

Se define como el cociente entre el número de pares que corresponde a la distribución observada, sobre el número de pares posibles que corresponde a la situación de máxima dispersión (igual distribución de casos entre las categorías). Por lo tanto; el índice varía entre 0 y 1.

$$0 \leq \mathbf{ID} \leq 1$$

Donde:

ID = 1 en la situación de máxima dispersión (o mínima concentración),

ID = 0 en la situación de dispersión nula (o total concentración).

Si consideramos como distribución observada una de las que en el ejemplo hemos llamado *situación intermedia* (intermedia 2), el índice resulta:

$$\mathbf{ID} = \frac{\text{nº pares observados}}{\text{nº pares posibles en situación de Máx. Dispersión}} = \frac{8}{9} = 0,89 \quad \text{u} \quad 89\%$$

Cuando el número de categorías y/u observaciones es relativamente grande, la determinación del número de pares posibles y de pares observados se vuelve dificultoso. En estos casos el **ID** se determina mediante la siguiente fórmula :

$$ID=\frac{k(n^2-\sum f_i^2)}{n^2(k-1)}$$

donde:

k : número de categorías de la variable

n : total de casos

fi : cantidad de observaciones o frec. Abs. en la categoría i-ésima.

Veamos la utilidad de este índice para comparar la heterogeneidad del motivo de la búsqueda de trabajo entre los hombres y las mujeres.

Motivo de la búsqueda de trabajo por sexo - Posadas-1986.

Motivo de Búsqueda	**Varones**	**Mujeres**
Completar Ingreso Familiar Básico	1140	262
Ampliar Ingreso Familiar Básico	452	490
Otros Motivos	578	702
Total	**2.170**	**1.454**

Fuente: EPH, mayo 1986.

Para la desviación de los varones el índice resulta:

$$ID=\frac{3[(2170)^2-(452^2+1140^2+578^2)}{2170^2(3-1)}=0,91$$

En el caso de las mujeres será:

$$ID=\frac{3[(1454)^2-(262^2+490^2+702^2)}{1454^2(3-1)}=0,93$$

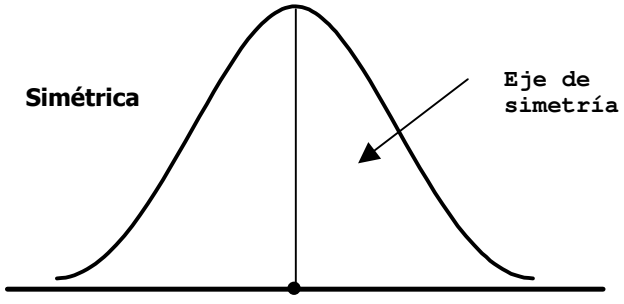
Ambos grupos presentan una alta dispersión (ID cercano a 1). Dado que el ID de las mujeres es mayor, *"las mujeres son ligeramente más heterogéneas que los hombres en cuanto al motivo por el que buscan trabajo".*

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.

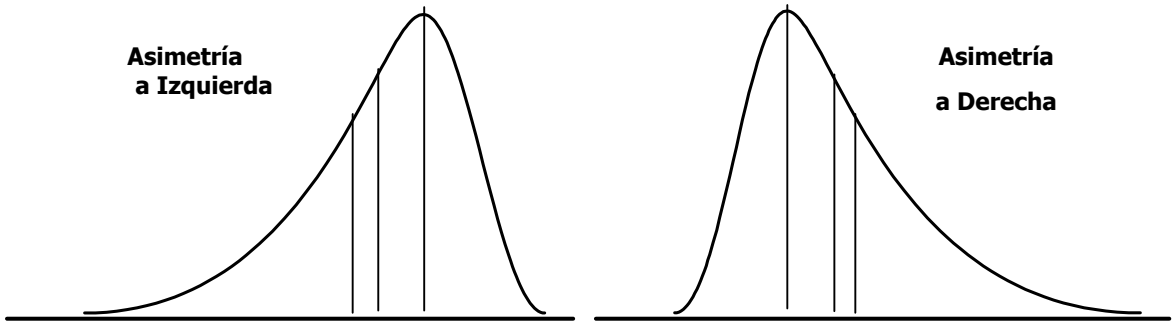
3. ¿Cómo Medir la Asimetría?

Como señaláramos oportunamente la "silueta" de la forma de la distribución (polígono de frecuencias) nos da una idea acerca de la simetría del conjunto de datos. Así teníamos que, en la situación de simetría, cada mitad de la curva es una imagen espejada de la otra mitad y la recta que hace de "espejo" (eje de simetría) es la que pasa por las medidas de tendencia central (media, mediana y modo, que coinciden en el mismo valor).



$\overline{x}$=Ma=Mo

A medida que la distribución se hace más asimétrica hacia uno u otro lado (derecha e izquierda), las medidas de tendencia central tienden a alejarse unas de otras, siendo la media -por estar afectada por los valores extremos- la que más se desplaza hacia la cola de la distribución (ver gráficos siguientes).



$\bar{x} < Ma < Mo$ $\qquad$ $Mo < Ma < \bar{x}$

Vemos en los Gráficos que, en el caso de una asimetría a la izquierda, la media es menor que la mediana y esta a su vez, menor que el modo. Inversamente, en la asimetría a derecha será el modo asume el menor valor y la media la mayor de las tres medidas. Se puede ver además que la mediana, siempre toma un valor intermedio entre las otras dos medidas, ubicándose más próxima a la media [6].

A medida que la **asimetría crece** en una u otra dirección, también las **distancias entre la media y el modo, y la media y la mediana, crecen**. En consecuencia, podemos utilizar estas diferencias ($\bar{x}$- Mo , o $\bar{x}$- Ma) como **medidas absolutas de la asimetría de una distribución**. Además se puede ver que si la asimetría es a la izquierda, $\bar{x}$- Mo dará un valor negativo, en tanto que si la asimetría es a la derecha esta diferencia será positiva.

En síntesis:

$$\bar{x}\text{-Mo} = 0 \Rightarrow \textbf{Simetría}$$

$\bar{x}$-Mo < 0 $\Rightarrow$ **Asimetría negativa** $\qquad$ $\bar{x}$-Mo > 0 $\Rightarrow$ **Asimetría positiva**

Además, cuanto mayor sea el valor absoluto de la diferencia, mayor será el grado de asimetría de la distribución

$$\text{A mayor } |\bar{x}\text{-Mo}| \quad \Rightarrow \quad \text{mayor asimetría}$$

Para poder **comparar la asimetría** de distribuciones de variables medidas en distintas escalas o presentadas para valores con distinta magnitud, la solución es **construir medidas relativas** de asimetría.

3.1. Coeficiente de asimetría de Pearson

Una de las medidas de asimetría más difundidas, es el **Coeficiente de Asimetría de Pearson** que calcula esa diferencia en cantidad de desvíos estándar.

Coeficiente de Asimetría de Pearson **(CAP)**

Se define como:

$$CAP = \frac{\bar{x} - Mo}{\sigma}$$

[6] En casos de asimetría moderada, la mediana se ubica -próxima a la media- a un tercio de la distancia entre la media y el modo.

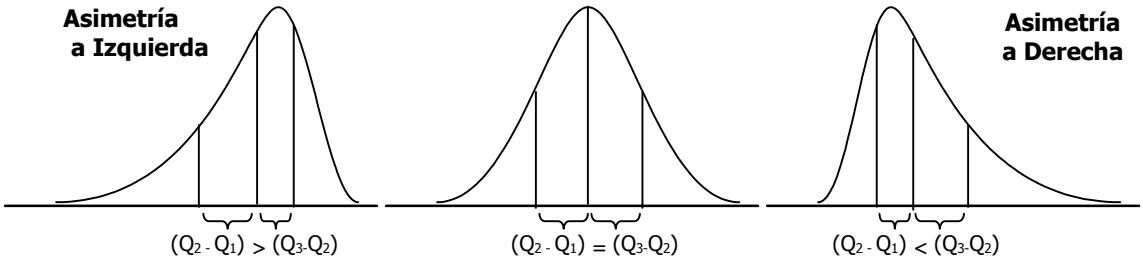
Comentarios

- La **magnitud absoluta** del coeficiente indica la **"cantidad de desvíos estándar"** a los que se encuentra la media del modo.
- Se lo puede expresar en porcentaje, multiplicando por 100 el resultado de la expresión anterior.
- Si el coeficiente es ***igual a cero***, estamos en una situación de ***simetría perfecta***.
- En situaciones de ***asimetría***, el coeficiente puede tomar valores positivos o negativos:
  - Los valores ***positivos*** están indicando una **asimetría a la *derecha***.
  - Los valores ***negativos*** indican una **asimetría a la *izquierda***.
- En términos teóricos, este coeficiente puede tomar valores que **varían entre –3 y +3**.

3.2. Coeficiente intercuartílico de Bowley

Una medida alternativa del grado de asimetría se puede plantear a partir de las distancias que se observan entre los cuartiles. En una situación de simetría los cuartiles 1 y 3 estarán equidistantes de la mediana. Es decir: $Q_3-Q_2 = Q_2-Q_1$

Ahora bien, si la ***distribución es asimétrica, estas distancias no serán iguales*** y variarán con el grado de asimetría; en consecuencia, las diferencias entre estas distancias pueden usarse como base para medir la asimetría de una distribución.



Tomando en cuenta esta característica de las distancias intercuartílicas, Bowley propone una medida relativa que expresa estas diferencias en términos del recorrido intercuartílico.

Coeficiente intercuartílico de Bowley (CAB)

Se define como:

$$CAB=\frac{(Q_3-Q_2)-(Q_2-Q_1)}{Q_3-Q_1}$$

De esta expresión, se deduce otra más sencilla para el cálculo manual.

$$CAB=\frac{Q_3+Q_1-2.Q_2}{Q_3-Q_1}$$

Comentarios:

- En situaciones de *asimetría*, el coeficiente puede tomar valores positivos o negativos:
  - Los valores *positivos* están indicando una asimetría a la *derecha*.
  - Los valores *negativos* indican una asimetría a la *izquierda*.
- En términos teóricos este coeficiente puede tomar valores que ***varían entre –1 y +1***.
- Según Bowley:
  - un valor de 0,1 (o –0,1) puede considerarse una ***asimetría moderada;***
  - un valor de 0,3 (o –0,3) puede considerarse como una ***marcada asimetría***.
- El coeficiente es *igual a cero*, en una situación de *simetría perfecta*.
- El coeficiente será 1 (o –1) cuando el Q1 (o Q3) coincida con la mediana.

Como parte de un estudio de medición de audiencia radial, se llevó a cabo una encuesta a 150 hogares de la ciudad para medir el tiempo de escucha de dos radios locales, entre las 16 y las 19 horas. Los resultados de esta observación se presentan en las tablas siguientes:

FM Guaraní

Tiempo de escucha (minutos)	Hogares (nº)
0 – 15	14
15 – 30	18
30- 45	20
45 – 60	25
60 – 75	45
75 – 90	18
90- 105	7
105 – 120	3
TOTAL	**150**

FM Acuario

Tiempo de escucha (minutos)	Hogares (nº)
0 – 15	3
15 – 30	45
30- 45	25
45 – 60	20
60 – 75	18
75 – 90	18
90- 105	14
105 – 120	7
TOTAL	**150**

MEDIDA	FM Guaraní	FM Acuario
$\overline{x}$	54,1 min	52,5 min
Ma	59,1 min	46,9 min
Mo	66,3 min	28,4 min
Q1	34,1 min	26,5 min
Q3	71,8 min	76,3 min
σ	25,8 min	28,9 min

El promedio de escucha en ambas radios es similar, aunque es de destacar que la mitad de los oyentes de radio Guaraní escuchan aproximadamente una hora o menos en esa franja horaria, mientras que la mitad de la audiencia de FM Acuario no excede los 47 minutos. Se destaca la diferencia en los tiempos más frecuentes de escucha (66 min. en Guaraní, y 28 min. en Acuario).

La heterogeneidad de los tiempos de audiencia es levemente mayor en FM Acuario (CVg = 0,48 y CVa = 0,55). A su vez, la distribución de los tiempos de escucha en FM Guaraní tienden a concentrarse en los valores más altos, mientras que los de FM Acuario en los valores más bajos; esto se manifiesta en los coeficientes de asimetría (negativo para el primer caso y positivo en el segundo). Además, es mayor el grado de asimetría en FM Acuario (0,83 veces el desvío estándar).

$$CAPg=\frac{54,1-66,3}{25,8}=-0,47 \qquad CAPa=\frac{52,5-28,4}{28,9}=0,83$$

Si ***analizamos la asimetría en el 50% central*** de los tiempos de escucha de ambas radios, se aprecia que en el caso de FM Guaraní es marcada la asimetría a izquierda en el grupo central, en tanto que en FM Acuario es moderada y a derecha.

$$CABg=\frac{(71,8-59,1)-(59,1-34,1)}{71,8-34,1}=-0,33 \qquad CABa=\frac{(76,3-46,9)-(46,9-26,5)}{76,3-26,5}=0,18$$

IMPORTANTE

Las diferencias entre el **coeficiente de Pearson y el de Bowley** están expresando con claridad que, aun cuando ambos miden asimetría, lo hacen sobre la base de criterios diferentes: el primero mide la asimetría de toda la distribución, mientras el segundo se refiere únicamente a los datos centrales. En consecuencia **aportan información complementaria** sobre esta característica de la distribución.

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***.**

4. ¿Qué Hemos Visto? (*)

En esta unidad hemos avanzado en la descripción de la forma de una distribución, presentando herramientas que nos permiten medir dos características centrales: ***variabilidad y asimetría***. Estas medidas **complementan las medidas resumen** presentadas en el Capítulo anterior.

Así entonces, hemos presentado medidas de dispersión para variables numéricas que se construyen sobre la base de diferentes **criterios: rango o campo de variación de los datos, y distancia de las observaciones a una medida de tendencia central** que se toma como referencia. Surgen entonces una serie de **medidas que expresan la cantidad de variabilidad en términos absolutos***.*

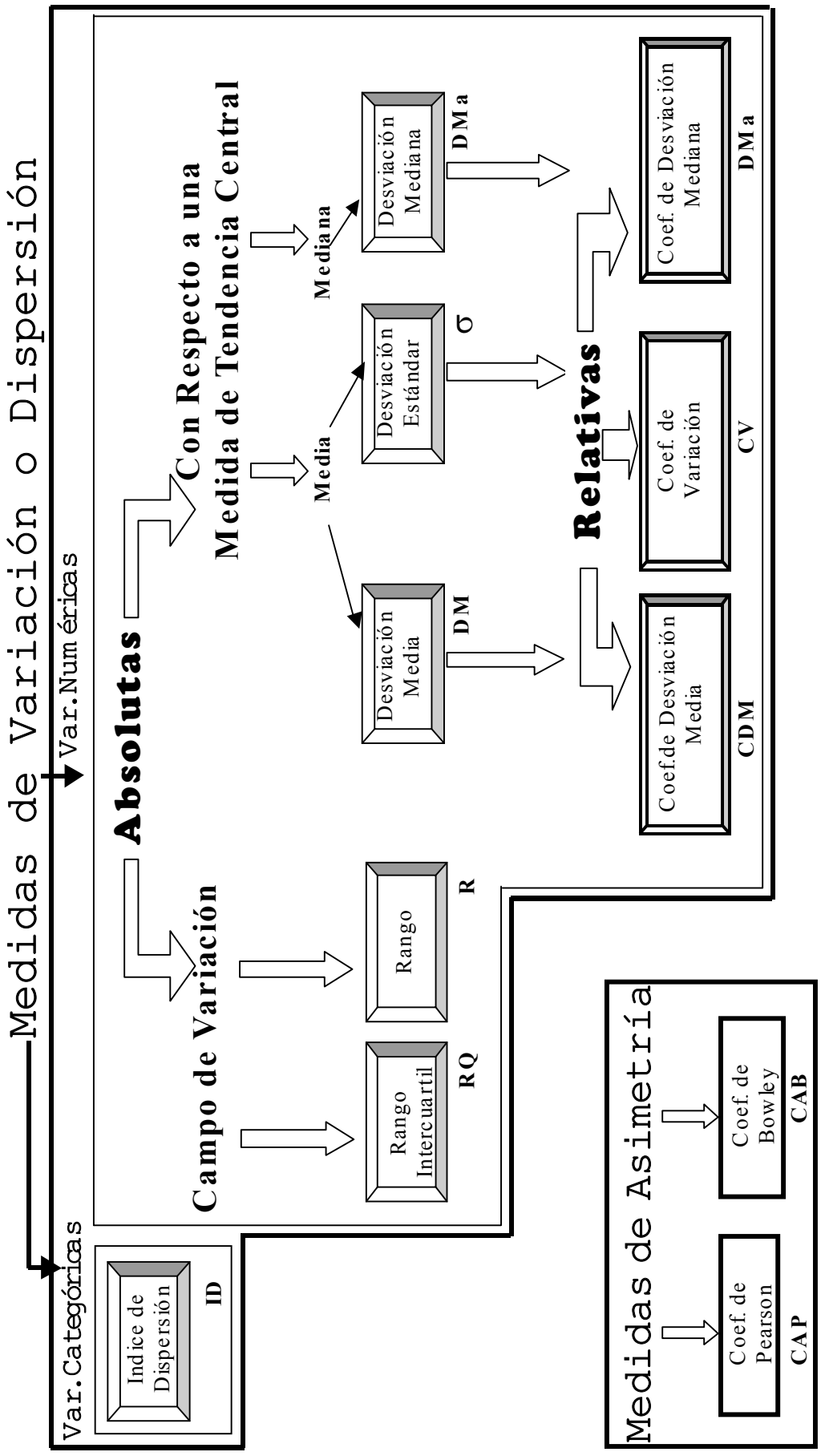
Para resolver cuestiones de **comparabilidad** de diferentes distribuciones, presentamos además **medidas relativas de dispersión**, transformando las principales medidas absolutas.

También, para medir la dispersión de **variables categóricas**, propusimos un **Índice de dispersión**.

Finalmente, presentamos medidas que valoran y permiten comparar el **grado de asimetría** de distintas distribuciones.

En todos los casos, se analizó la variabilidad o la asimetría, en ejemplos que ayuden a la interpretación y comunicación de estas herramientas de análisis, destacando su complementariedad con otras herramientas de análisis.

(*) Ver esquema en la página siguiente


Medidas de Variación o Dispersión
Var.Numéricas
Var.Categóricas
Indice de Dispersión
ID
Absolutas
Campo de Variación
Rango
R
Rango Intercuartil
RQ
Con Respecto a una Medida de Tendencia Central
Media
Desviación Media
DM
Desviación Estándar
σ
Mediana
Desviación Mediana
DMa
Relativas
Coef.de Desviación Media
CDM
Coef. de Variación
CV
Coef. de Desviación Mediana
DMa
Medidas de Asimetría
Coef. de Pearson
CAP
Coef. de Bowley
CAB

Bibliografía

BARBANCHO, A. (1978): *Estadística Elemental Moderna*. Ed. Ariel, Barcelona, España. Páginas: 145-146.

BLALOCK, H. M. (1986): *Estadística Social*, México, FCE. Páginas: 90 a 102.

SHAO, S. (1967): *Estadística Para Economistas y Administradores de Empresas*. Herrero Hermanos S.A., México. Páginas: 218 a 237.

UNIVERSIDAD NACIONAL DE CÓRDOBA (1993): *Estadística aplicada a la Investigación. Curso a distancia*. Fac. de Cs. Económicas, Córdoba, 1993. Módulo IV. Páginas: 3-16.

Conceptos Centrales

- Variabilidad / Dispersión.
- Necesidad de medir la variabilidad.
- Criterios para construir medidas absolutas de dispersión para variables numéricas
- Necesidad de utilizar medidas relativas de dispersión o variabilidad.
- El concepto de dispersión para variables categóricas y la medición asociada.
- Concepto de Asimetría y criterios para su medición.

Habilidades

- Seleccionar y obtener las medidas de variabilidad más apropiadas a una situación de trabajo.
- Interpretar las diferentes medidas en términos del problema.
- Comparar la variabilidad de diferentes distribuciones.
- Seleccionar y obtener medidas de asimetría.
- Interpretar las diferentes medidas de asimetría.
- Describir la forma de una distribución integrando las diferentes medidas de resumen conocidas.
- Comunicar los resultados del análisis.