CAPITULO 3

ANALISIS UNIVARIADO

Conocidos los personajes centrales del juego cuantitativo, casos, valores y *variables* (que son todos los valores de los casos, vistos en columnas) nos proponemos en el presente capítulo conocer lo básico del análisis de las variables, tomadas cada una por separado: este trabajo se denomina análisis *uni*variado. Cuando se trabaja al mismo tiempo con dos variables, tenemos análisis *bi*variado; con tres, análisis *tri*variado (o estratificado); cuando son más de tres las variables que se analizan en conjunto, se acostumbra hablar de análisis *multi*variado. Estos tipos de análisis tienen un propósito central: conocer la *variación básica* de las variables en cuestión y dar cuenta razonable de los posibles errores que afectan esa variación, es decir de la *variación espúrea*.

Los tipos de variación que hay en las variables de una base de datos

El trabajo de análisis de variables se inicia conociendo el **comportamiento** empírico de la *variación* de los valores que hay en la columna. En el capítulo 2 introdujimos el tema de la variación, ahora volvemos sobre ella para hacer un par de distinciones que son muy importantes. Una variable es *variable* porque no es constante en sus valores legales, como explicamos en el capítulo 2. Esa variación de los valores, que tiene un *rango*, se denomina *básica* o fundamental, para distinguirla de otro tipo de variación que se denomina *de error*, el cual nace de factores diferentes a los que causan la variación básica. Por ejemplo: si en una sala hay 200 personas de las cuales 150 son hombres y 50 mujeres (en porcentaje: 75% y 15%), la *variación básica* es la que nace de los factores que hacen que haya en efecto 150 hombres y 50 mujeres. Puede suceder, sin embargo, que en mi base de datos sobre esa sala aparezcan apenas 140 hombres (70%) por factores extraños a la variación básica, por ejemplo, por *errores* de identificación de los varones, de conteo, o de digitación de los datos. Esta *variación espúrea*, debida a errores, es posible y se debe evitar o controlar para reducirla al máximo, cuando no se puede anular. Una forma muy especial de variación espúrea (es decir, distinta de la básica o fundamental), que será objeto de trabajo cuando veamos la inferencia estadística, es la debida al *error de muestreo*. Sin entrar en detalles podemos anticipar lo siguiente: cuando se trabaja con una porción de la totalidad de casos, por ejemplo, con 1.000 hogares de los 30.000 que tiene una región, se dice que uno trabaja con una *muestra*. Obviamente, el interés es *inferir a la población total*, es decir, con base en la muestra hablar de la totalidad de los hogares. En otras palabras nos interesa es la variación básica de las variables de la población, no de la muestra. Como las dos variaciones no siempre coinciden (casi nunca coinciden), se dice que en el trabajo con muestras uno introduce un error de muestreo, que se debe controlar. En resumen: la variación de las variables que uno tiene en una base de datos se divide en la variación básica, y la variación por errores, de los cuales uno muy importante en estadística es el error de muestreo. Este resumen se muestra en la figura 3.1

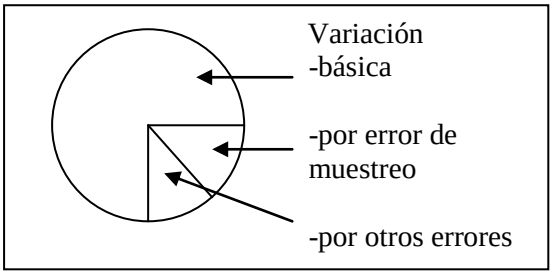

Figura 3.1: Tipos de variación que se encuentran en una base de datos (de un estudio muestral)

Primera aproximación a los datos: el listado de variables y valores,

Lo primero que un investigador hace con una base de datos es mirarla en su formato final. Parece una bobada decirlo, pero se dan casos en que se inicia el trabajo más complejo y no se ha cumplido con lo elemental: ver los datos en su forma final, por columnas y por líneas. Esta mirada ayuda a que el investigador se familiarice con sus datos y detecte a primera vistas irregularidades que se puedan presentar, porque casi siempre aparecen. Para hacer esto uno puede mirar la base de datos en su formato sin procesar, es decir en la base original (por ejemplo en la hoja electrónica en que se metieron), o en la base especial a que se convierten cuando uno utiliza un paquete estadístico y *entra* allí sus datos (es decir, abre desde el paquete, una base para trabajo con el mismo). Si ya la ha introducido a su paquete, puede pedir la lista de sus datos, y mirar variable por variable, de arriba abajo. Aconsejamos este procedimiento preliminar porque aparte de la familiarización con los datos en su forma primaria, cosa que ya de sí es importante, si uno tiene "buen ojo" puede detectar cosas interesantes.

(T y AL: dar un ejemplo de lista de variables, tomado p.e. de Epiinfo).

Frecuencias absolutas y relativas

Se denomina *distribución de frecuencia* el resultado del conteo de los casos por cada valor de una variable: es decir, se observa cómo se *distribuye* el total de los casos en las diferentes casillas discretas correspondientes a cada valor de la variable. En un ejemplo anterior, de la sala con 200 personas, la distribución de frecuencias de la variable "Género" es 150 para el valor "1" (hombres) y 50 para el valor "2" (mujeres). Esta es la *frecuencia absoluta*, con total 200. Cuando se utiliza un total constante, para facilitar comparación, por ejemplo 100, se habla de *frecuencia relativa* (en este caso "relativa a 100" o "porcentaje"), que es 75% de hombres y 50 de mujeres, es como decir, "de 100, 75 son hombres, y 50 mujeres".

El primer proceso que uno hace con una base de datos, una vez que la ha construido y mirado al detalle las listas de valores, es sacar sus frecuencias, absolutas y relativas. Utilicemos el ejemplo, tomado de una encuesta sobre prácticas y creencias sexuales realizada en Colombia por el Seguro Social en 1993. [1] Había una pregunta sobre el número de parejas sexuales que el encuestado había tenido en el año anterior, que se convirtió en la variable "Parejas sexuales año anterior" cuyas frecuencias absolutas y relativas aparecen en la tabla 3.1. Mirémosla en detalle y luego hagámosle algunas observaciones.

Tabla 3.1. Ejemplo de Frecuencias de la Variable "parejas sexuales en año anterior"

[1] Fuente: Base de datos de la Encuesta ISS/Colombia/1993, "Conocimientos, actitudes y prácticas referentes a ETS/SIDA en Colombia". Ver E. Sevilla, J. Olaya y K. Feliciano, Dueños de sí y de sus deseos: la conducta sexual de los colombianos y su vulnerabilidad al VIH. Cali: Universidad del Valle, 1994.

Valores	***Frec. Absoluta***	***Frec. Relativa base 1 (p)***	***Frec. Relativa base 100***	***Frec. Acumulada base 100***
0	*1393*	*0.116*	*12.7*	*12.7*
1	*7576*	*0.693*	*69.3*	*82.0*
2 a 4	*1463*	*0.134*	*13.4*	*95.4*
5 y más	*507*	*0.046*	*4.6*	***100.0***
Total	***10939***	***1.000***	***100.0***	

Observaciones a la tabla 3.1

a. Como aparece en la tabla es una variable *categórica ordinal*, cuyos valores se organizan en cuatro categorías discretas que van de 0 a 5+. La base de datos original estaba en números reales, que iban de 1 a n, es decir se trataba de una *variable intervalo*. Sin embargo, para el análisis se optó por recodificarla, volviéndola variable categórica ordinal, con cuatro valores crecientes, "0", "1", "2 a 4", y "5 y más". ¿Por qué estas cuatro categorías, y no otras, o más? La respuesta depende del comportamiento de la variable intervalo original y de otros criterios prácticos o teóricos: en este caso el investigador decidió, mirando la frecuencia original de la variable intervalo (valores 1, 2, 3, 4, 5…) que era mejor, para propósitos de análisis y conclusiones, recodificarla en categórica ordinal con esos cuatro valores. Uno de los criterios que se utilizó para recodificar así fue la comparabilidad con otras encuestas preexistentes que ya habían recodificado en esos mismos cuatro valores (ver párrafo siguiente).

b. La frecuencia absoluta tiene un total (o base) contingente, no constante, que en este caso es 10939, pero que pudo ser otro, por ejemplo 8040 o 20458. Supongamos que una encuesta similar para EEUU, de 3155 personas produjo para el valor para "5 y más" una frecuencia absoluta de 202 personas Si, haciendo un **análisis univariado con dos muestras**[2], **comparamos** la variable para el valor "5 y más" en Colombia y EEUU, tenemos cierta dificultad en responder a la pregunta ¿En dónde se toman más libertad sexual para escoger parejas: en Colombia donde hay 507 en tal categoría o en EEUU que apenas muestra 102?. La respuesta cruda, no ajustada a una base común de comparación, es que en Colombia, puesto que hay 507 casos y en EEUU sólo 102. Pero es obvio que esta respuesta tiene una falacia escondida, pues las cifras no son comparables dado que se refieren a totales diferentes. Es preciso buscar una *base comparativa única*, es decir, **relativizar** los datos con respecto a una misma base de comparación: de allí surgen las **frecuencias relativas**.

c. La frecuencia relativa se hace con una **base fija** o común, es la unidad (1.0). Se obtiene dividiendo la frecuencia absoluta (507 en Colombia, 202 en EEUU) por el total absoluto respectivo (10939 en Colombia, 3155 en EEUU). Al hacer estas divisiones resultan valores decimales, denominados **proporciones**, concretamente 507/10939=0.046 para Colombia, y 106/3155=0.064 para EEUU. Ahora sí se pueden comparar los dos países con respecto a esta variable: Colombia con 46 milésimas tiene una proporción inferior a EEUU que tiene más, exactamente 64 milésimas. Estas proporciones se leen así: "de 1000 personas en Colombia que respondieron la encuesta 46 tuvieron

[2] Obsérvese que pudimos decir: análisis bivariado, pues introdujimos implícitamente la variable PAIS. Mantenemos, sin embargo, el concepto de univariado pues la variable PAIS no ha sido explicitada ni sistematizada. Para ello hemos debido crearla, dándole valores explícitos.

5 y más parejas sexuales el año anterior; y 64 en EEUU". La proporción es, por tanto, el resultado de **relativizar** una frecuencia con referencia a la *base 1*, pues las frecuencias relativas así generadas deben sumar 1.000, con la aproximación decimal que uno considere conveniente (aquí en milésimas). En la tabla esto se puede comprobar en la tercera columna, cuyas frecuencias relativas de base 1 suma exactamente 1.000.

c. Las proporciones originales de base uno (0.032 y 0.046 en nuestro ejemplo) son un poco incómodas para leer, sobre todo en comunicaciones destinadas al gran público. Por ello se acostumbra usar **el porcentaje**, multiplicando por cien las cifras decimales, es decir, corriendo hacia la izquierda 2 cifras el punto decimal. Así se generó la columna 4 de la tabla. Las cifras comparativas del ejemplo de Colombia y EEUU quedan entonces en y 4.6% y 6.4%, que son más legibles. A partir de la base 1 podemos multiplicar nuestras proporciones por 10, por 100 (%), por 1000 (o/oo), etc., según la necesidad y el margen de maniobra que nos permitan los datos.

☝ Un consejo importante sobre el uso de las cifras decimales o de precisión que se dan en un informe. "Ni tanto que queme el santo y tan poco que no lo alumbre". Un recurso falaz es el de dar idea de alta precisión aumentando las cifras decimales, por ejemplo 4.63% y 3.23%. El límite justo de decimales lo impone la naturaleza de los datos y la sindéresis del investigador frente a la precisión requerida. En el ejemplo, 3% y 4% ocultan información importante, puesto que la variación entre los dos países, para este valor de la variable, es relativamente pequeña y vale la pena trabajar con décimas de porcentaje. Dos cifras decimales (centésimas) no añaden información y estorban. Por tanto, una cifra decimal en el porcentaje es lo justo.

d. La base relativa 100 (porcentual) es la más común para eventos con frecuencias corrientes. Por ello los paquetes estadísticos la producen automáticamente (lo mismo que los noticieros). Cuando los eventos son raros hay que utilizar otras bases. Por ejemplo, los homicidios son muy raros en la mayoría de sociedades, por ello la tasa comparativa internacional es "por 100.000". (T&AL ampliar el ejemplo).

Ejemplo: En Colombia, donde estos eventos de homicidio no son tan raros los especialistas usan frecuentemente la tasa por 10.000 pero debe utilizarse "por 100000" si uno desea hacer comparaciones internacionales, pues en otros países, con que se compara Colombia, los homicidios son más raros. El traslado a otras bases relativas, 10, 100, 1000, etc. depende, por tanto, de la frecuencia o rareza del evento y de la necesidad de comparación. Las tasas o proporciones entonces vienen expresadas en "por ciento", "por mil", "por diezmil", "por cienmil" etc. todas ellas construidas a partir de la base unitaria o proporción simple. Ejemplo: si encontramos 13 infectados de VIH en una muestra de 3500 sueros sanguíneos analizados por la Secretaría de Salud de Cali tenemos la proporción simple de 3/3500 = 0.0037. Como es usual en estudios de VIH/SIDA utilizar la base "por mil"para las infecciones, multiplicamos la p hallada por 1000 y podemos decir que en Cali, a partir de la muestra utilizada, encontramos una prevalencia de VIH positivo de 3.7 por mil. Obsérvese que el punto de partida siempre es la proporción simple de base 1.

e. La tabla 3.1 trae la columna de **frecuencias acumuladas**, las cuales pueden ser absolutas, o relativas (éstas de diferente base, 1, 10, 100...). La frecuencia acumulada es muy útil para ciertos análisis en que es importante decir (en escalas ordinales) "de este punto para abajo hay tantos casos

Por ejemplo, en la tabla 3.1, por la frecuencia acumulada porcentual podemos decir que las personas que reconocieron menos de 5 parejas en el año anterior corresponden al 95.4%.

f. Recordemos que como las variables continuas pueden tratarse como discretas, de ellas también se pueden sacar frecuencias, aunque admiten operaciones lógicas y matemáticas más complejas que veremos más adelante. De todos modos es aconsejable en un primer examen de la base de datos sacar frecuencias de todas las variables, categóricas y continuas. Mirar estas frecuencias, absolutas y relativas, es muy conveniente para perfeccionar la familiaridad que un investigador debe tener con sus variables de trabajo.

i. Las frecuencias relativas que sirven muy bien para trabajar con rigor y finura determinadas comparaciones cuando los valores absolutos contienen eventuales falacias, como es el caso arriba reportado de la frecuencias de ¨5 y más parejas sexuales", en EEUU y Colombia, falacia que se repite el ejemplo siguiente:

Ejemplo adicional, ficticio: Se hace una comparación entre los "proyectos aprobados" por Colciencias a las Universidades Nacional, de Antioquia, y del Valle, respectivamente: 103, 53, 45 para determinar qué universidad "investiga más". Vistos así, en valores absolutos, el orden decreciente es: Unal, Udea, y Univalle. Pero sacando tasas (usando como denominador el número de profesores de cada universidad) tenemos: 103/2300=0.045 , 53/1200=0.044 y 45/1000=0.045, es decir muy pocas diferencias, que se miden el milésimas o tasas por mil.

Los resúmenes de las variables: segundo paso del análisis univariado

Una vez estudiadas las distribuciones de frecuencia, se pasa a resumir las variables para su uso comparativo. Se habla de **resúmenes** de las variables porque distribución de frecuencias es una tabla en donde se observa el resultado del **conteo** de ocurrencias de casos para cada valor en la escala de la variable, pero tal tabla se hace inmanejable para efectos prácticos. Los resúmenes de la distribución permiten trabajar cada variable de manera elegante y práctica con **una sola cifra** numérica que habla por el conjunto de la variable. Los resúmenes son de tres clases: las **proporciones** o **tasas**, las medidas de **tendencia central**, y las medidas de **dispersión**. De las proporciones o frecuencias relativas ya hablamos un poco. Se llama medida de tendencia central a aquel resumen de la variable que expresa la tendencia de los casos a agruparse alrededor de un valor que se considera es el mejor representante de la variable; y medidas de dispersión las que muestran la tendencia contraria, a dispersarse o alejarse de ese punto central. Veamos cómo se generan y usan estos resúmenes en el caso de las variables categóricas y de las continuas, porque aquí ya comienzan a darse diferencias.

Para variables categóricas

Proporción o tasa. Es muy frecuente concentrarse en el análisis de una sola casilla, para propósitos de comparación. Lo hicimos arriba comparando la frecuencia de "5 y más parejas sexuales" en Colombia y EEUU. En tal caso, a la proporción de interés, que se denomina **p,** se le da un complemento, constituido por el resto de valores de la variable; en el ejemplo de la tabla 3.1, los valores "0", "1", y "2 a 4". Este resto se denomina **q**. Como se sabe que las frecuencias relativas de

base 1 suman 1, se aplica la ecuación **p+q=1**, o cualquiera de sus transformaciones algebráicas, de las cuales la más usual es **p = 1-q**, o **q = 1-p.** (T&AL: dar ejemplo).

Un nombre muy usado para las proporciones así generado es el de **tasa**. Hay distinciones adicionales en las **tasas** que se verán luego, de las cuales las más comunes son la **prevalencia** y la **incidencia** sobre las que decimos una palabra aquí. Brevemente, la prevalencia indica la proporción en un punto del tiempo de determinados **eventos o condiciones** con respecto a una base poblacional o muestral, p. e. la prevalencia de enfermos de sida en Cali a mayo de 1995 se obtiene dividiendo el número de enfermos detectados *hasta la fecha* (es decir el acumulado) por la población de la ciudad. La incidencia, en cambio, indica los **nuevos eventos** registrados durante **un período** determinado, p. e., los *nuevos* casos detectados de sida en la ciudad en lo que va corrido de 1995. (T&AL: perfeccionar el ejemplo).

La proporción o tasa no es en realidad un resumen completo de la variable, y ciertamente no es una medida de tendencia central ni de dispersión. Como reduce la variable a una sola cifra de interés y es muy usada como resumen parcial de una variable, la tratamos aquí como primera forma de resumen en las variables categóricas.

La moda. Otra medida común en las variables categóricas es la **moda** o **valor típico.** Es una medida resumen que nos indica el valor más común, o los valores más comunes (pues puede haber varias modas), que hay en una distribución de frecuencias, sin atender a los valores extremos. Por cuanto la moda señala el valor más común puede denominarse medida de tendencia central en las variables categóricas. Sin embargo, una distribución de frecuencias puede tener varios picos o modas. Tomando la variable edad en el ejemplo de la Tabla 3.2, (una variable continua tratada aquí como categórica) se podría decir que las edades de 20, y de 28, son la moda, pues son las de mayor frecuencia, absoluta o relativa. Algunos estadísticos dicen que también se puede sacar la moda de las variables continuas (en el caso de que edad se trabaje como continua, por ejemplo): esa moda sería el puntos, o los puntos, alrededor de los cuales se concentra el mayor número de casos. Pero como las variables continuas tienen mejores resúmenes que la moda, esta aplicación poco se usa.

La moda es poco usada en sociología, aunque podría ser útil en ciertos casos cuando uno quiere dar cierto sabor "cuantitativo"a conclusiones referentes a los atributos predominantes en cierto conjunto. Hubo en antropología cierta escuela que propuso la "personalidad modal" en una cultura, teniendo como idea de trasfondo la frecuencia modal de ciertas estructuras de personalidad individual.

Una nota marginal: es curiosa la relación entre el significado estadístico de la moda como resumen de variable y el significado sociológico de la moda como rasgo de conducta colectiva: alguien lanza una moda, p. e. de corte de cabello masculino y, mediante un proceso digno de ser estudiado más a fondo, rápidamente se expande su uso de tal modo que la moda (social) se convierte en moda (estadística). De este modo se origina otra moda. Curioso, ¿no es cierto?.

El rango y los percentiles. A partir de la distribución de frecuencias se pueden establecer rápidamente unas primeras **medidas de dispersión** de la variable categórica (o continua tratada como categórica) **el rango** y, sobre todo, los **percentiles**.

El **rango** es la distancia *empírica* entre el valor máximo y mínimo que ofrece la distribución de la variable. En el ejemplo de la Tabla 3.1 la distancia empírica, es decir la que ofrece la variable en cuestión, el rango de "Edad" es 16-47. Este rango, correspondiente en el ejemplo a una muestra de trabajadoras sexuales del centro Cali, se podría comparar con el rango de otro grupo del mismo tipo de personas perteneciente a otro sector social de la ciudad, más "exclusivo" y más controlado por la policía; podría ser un rango de 18-35, dado que la edad mínima para este tipo de trabajo es 18, y chicas con edades superiores a 35 ya no tendrían clientes.

Los **los percentiles** se generan con base en la frecuencia relativa acumulada. El rango de la variable se divide en 100 partes, de las cuales cada una es un percentil. Como no siempre hay casos en cada percentil de los 100 posibles, las tablas usualmente omiten los percentiles con frecuencia 0; además, las tablas de frecuencia rara vez presentan percentiles enteros: los dan en dedimales, como puede verse en la Tabla 3.1, que como dijimos corresponde a la edad de trabajadores sexuales del centro de Cali. El percentil 1 (cercano al 0,8%) corresponde aproximadamente a la edad 16 años, el percentil 25 a la edad de 21 años, y el percentil 80 cercano a la edad de 30 años. (Queda como ejercicio para el estudiante el pensar cuáles serían los mismos percentiles comparables en el caso de las trabajadores sexuales del sector exclusivo arriba mencionado, con rango de edad entre 18 y 35). Si uno desea obtener percentiles exactos basta que los pida al paquete estadístico que está usando. Así, por ejemplo, el SPSS produce para el caso de las chicas del centro de Cali los siguientes percentiles exactos: 1=…(T y AL, generarlos por favor).

Los **cuartiles, quintiles y deciles** son términos técnicos, comunes en algunos estudios, que en lenguaje común se traducirían por las cuartas, quintas y décimas partes de una distribución de frecuencias relativas. Se trata en rigor técnico de los percentiles correspondientes al total 100 dividido en 5, 4 y 10 partes. Se habla entonces, por ejemplo, del *primer cuartil* que corresponde al percentil 25 (porque 100/4=25), del *segundo cuartil* que corresponde al percentil 50, del *tercero* al percentil 75, y del *cuarto* al percentil 100. Lo mismo se hace con los quintiles y deciles.

Tabla 3.2. Ejemplo de tabla de frecuencias producido con SPSS

EDAD

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	16	2	,8	,8	,8
	17	5	1,9	1,9	2,7
	18	16	6,1	6,1	8,8
	19	11	4,2	4,2	13,0
	20	19	7,3	7,3	20,2
	21	14	5,3	5,3	25,6
	22	19	7,3	7,3	32,8
	23	17	6,5	6,5	39,3
	24	15	5,7	5,7	45,0
	25	16	6,1	6,1	51,1
	26	15	5,7	5,7	56,9
	27	17	6,5	6,5	63,4
	28	19	7,3	7,3	70,6
	29	15	5,7	5,7	76,3
	30	11	4,2	4,2	80,5
	31	11	4,2	4,2	84,7
	32	4	1,5	1,5	86,3
	33	6	2,3	2,3	88,5
	34	6	2,3	2,3	90,8
	35	6	2,3	2,3	93,1
	36	5	1,9	1,9	95,0
	37	4	1,5	1,5	96,6
	38	1	,4	,4	96,9
	39	2	,8	,8	97,7
	40	2	,8	,8	98,5
	41	1	,4	,4	98,9
	43	1	,4	,4	99,2
	44	1	,4	,4	99,6
	47	1	,4	,4	100,0

Tabla estándar de frecuencias producidas por SPSS. Se observa en primer lugar el **rango** de valores, que va de 16 a 47, es decir (47-16=) 31 años. La **frecuencia absoluta** nos indica el número absoluto de casos de hay en cada "cajón" de valor, ej. 5 en los 17 años. El *Percent* muestra la **frecuencia relativa** teniendo, en este caso, como referente total 100; el *Percent* y el *Valid percent* en este caso son iguales porque no hay valores nulos o *missing*. Esta frecuencia relativa se lee así, por ejemplo: de 100 casos un 1.9 están en los 17 años. El *Cumulative Percent* sirve, entre otras cosas, para marcar **los percentiles:** se ve que el percentil 25 (o cuartil 1, Q_1) está cercano a los 21 años; el segundo a los 25; y el tercero a los 29. Lo que quiere decir, por ejemplo y usando el Q_2, que la mitad de esta gente está en 25 años o por debajo de esa edad; y las tres cuartas partes en o por debajo de 28.

Para las variables continuas

La **media o promedio artimético** es una medida de **tendencia central** muy usada por lo fácil de su cálculo, por sus propiedades numéricas reales (resulta de, y admite operaciones aritméticas), y porque viene asociada a un complemento muy importante, la **desviación estándar** que es una medida de la **dispersión** de la variable. En este importante binomio, **media y su DE**, tenemos ordinariamente un excelente resumen de cualquier variable continua. Dada su importancia, nos vamos a detener un poco no tanto en los aspectos técnicos de su producción como en la interpretación de su significado. Hay otras **medias** (geométrica, armónica, acotada…) que poco se usan en sociología.

📖 Sugerencia: repase cuidadosamente su manual de estadística descriptiva sobre los temas de media, varianza, desviación estándar, percentiles, mediana y moda. Igualmente estudie la relación que se da entre media, mediana y moda en diferentes curvas de distribución, y entre media, varianza y desviación estándar. Tener claridad sobre estos conceptos es un requisito para la comprensión de lo que sigue.

La **media** aritmética expresa un valor resumen de la variable en cuestión teniendo en cuenta **todos** los valores numéricos de la distribución de la variable. Por ello, si entre estos valores hay casos extremos, hacia arriba o hacia abajo, la media puede resultar un resumen falaz. De hecho, con una media mal utilizada se puede embaucara más de uno.

Ejemplo: la media de los valores 1, 2, 4 y 93 es 25; igualmente, la de los valores 24, 25, 25, y 26 es 25. Desde luego, en el primer caso la media no es un buen resumen de la variable pues la dispersión es muy grande. Sí lo es en el segundo caso en que la dispersión de los valores alrededor de 25 es muy escasa. La desviación estándar, como veremos, mide esa dispersión y nos ayuda, a través del coeficiente de variación, a evitar trampas de falacia con el uso de la media.

La **desviación estandar** (DE) es el **necesario complemento** de la media, pues su tamaño relativo a la media (llamado coeficiente de variación, **CV**= DE/media*100), nos indica si la media es un buen resumen de la variable. La idea básica detrás de la DE es la siguiente: toda variable continua tiene dos dimensiones resumen importantes: (i) la tendencia central, que nos da una idea de "alrededor de qué valor" de mueven los casos, y que es expresada por la media; ejemplo: el peso de los alumnos del salón "gira"alrededor de 160 kilos; y (ii) la dispersión de los casos con referencia a ese punto central, que está expresado por la DE; ejemplo: **la distancia promedia de desviación** del peso de los alumnos del salón, con respecto a 160 kilos es de 20 kilos.

Tabla 2.3 Ejemplos de Medias, DS, CV(y Mn)
Cálculo con una variable "edad" de un grupo de estudiantes

A

Valor	17		19	20 20	21		**23**		25		27	28	30		**...**	
D	-6		-4	-3 -3			0		+2		+4	+5	+7		**....**	
d^2	36		16	9 9	4		0		4		16	25	49		**....**	

Media = 23.0; n=10; ∑d2=168; Var= 168/10=16.8; DE=√16.4=4.1;
CV=4.1/23=0.178; (Mediana=22)

B

Valor	17	18	19	20 20	21		**23**		25		27				...	40
Dif	-6	-5	-4	-3 -3			0		+2		+4				...	+17
dif2	36	25	16	9 9	4		0		4		16				...	289

Media = 23.0; n=10; ∑d2=408; Var= 408/10=40.8; DS=√40.8=6.4;
CV=6.4/23=0.278; (Mediana=20.5)

La generación de la DE es simple: se ordenan los valores de menor a mayor; se calcula la media; se calcula la distancia (d) en valores positivos o negativos que cada valor tiene con respecto a la media; se elevan esas distancias al cuadrado (d^2) para evitar sumas cero; se suman esos d^2 y se dividen por el número (n) de casos y se obtiene un valor medio --al cuadrado-- llamado **varianza**; se saca la raíz cuadrada de la varianza y se obtiene la anhelada **DE**. Así de simple. Las tabla 2.3 trae ejemplos. (De paso se incluye la mediana, para observar su comportamiento con respecto a la media).

8. El **coeficiente de variación (CV)** es una medida **relativa**, ordinariamente dada en %, que señala qué tan grande o pequeña es la DE con respecto a la media. Ejemplo, tomado de la tabla 2.3 : la media en los dos casos de edad de 10 alumnos de un curso es 23. Sin embargo, en A la DE es 4.1 años, que es el 18%, (*4.1/23=0.178*), de la media 23, es decir tiene un CV de 18%. En cambio, en B la DE es 6.4 años, y su CV es el 28%, (6.4/23=0.278). El origen de esta diferencia en dispersión puede observarse en la tabla: en A no hay valores extremos, pues el rango va de 17 años a 30, mientras en B el rango va de 17 años a 40. El CV es una medida muy sensible y útil, para dar un uso correcto y sobrio a las medias (tan profusamente utilizadas).

☝ Consejo práctico: se tiene una regla simple, que los sociólogos suelen contar al oído a sus amigos: cuando el CV de su media se mueve entre el 10 y el 20% puede obrar con tranquilidad; cuando supera ese rango hacia abajo o hacia arriba, tenga cuidado. Hacia abajo, Usted está cercano a una constante, es decir no tiene mucha variación en su variable. Hacia arriba, su media deja de representar bien su variable, la cual tiene demasiada variación (de pronto, por valores extremos: en tal caso, use la mediana). La moraleja es pues: calcule la media; calcule su DS; y calcule su CV; y aplique la regla del 10-20%. (Como los paquetes estadísticos producen siempre la DS, y desde luego la media, Ud. no tiene sino que comparar mentalmente su DS con el 10% de su media, y proceder en consecuencia.

Ejemplo: Su media de ingreso familiar es 200 mil pesos y su DS es 100 mil pesos. El 10% de 200 es 20 mil, y el rango aconsejado para la DS es 20-40 mil pesos. Como 100 mil está muy por encima de

40 mil, tenga cuidado --por lo menos-- al usar esa media. La alternativa que le queda es o mirar la variable para dar tratamiento especial a los *outliers* (casos extremos o atípicos) que le están introduciendo tanta variación, o utilizar la mediana. (O buscar otro indicador de nivel socioeconómico).

La **mediana** es un resumen **posicional**, **no afectada por valores extremos**, que es útil para dar una idea de la **tendencia central** de una variable **numérica real** en la que hay motivos de peso para no usar la **media** (con su DE). Los motivos para preferir una mediana a una media son los siguientes, solos o en combinación: (i) la escala no tiene intervalos equivalentes, es decir, se trata en rigor de una variable continua, pero ordinal no intervalo; ejemplo: las calificaciones académicas deberían trabajarse con medianas, pues la distancia entre un 2.0 y un 3.0 no es equivalente a la que hay entre 4.0 y 5.0, como bien puede decirlo cualquier estudiante que ha perdido una materia; (ii) la escala tiene intervalos o clases abiertas, p. e. en la variable edad "menores de 20", o "mayores de 90"; y (iii) aunque la variable es intervalo y tiene clases cerradas su dispersión o variación, medida por el coeficiente de variación, indica que la media no es un resumen representativo de la variable.

Valores extremos o atípicos

El último punto merece una ampliación: hay series de casos, que constituyen una variable, cuyos valores son muy desiguales porque tiene *outliers* o **valores extremos** que "dañan" cualquier análisis, introduciendo modificaciones fuertes a la media, que como se sabe es afectada por **todos** los valores de la serie. En tal caso el analista puede, o bien excluir los *outliers*, o bien utilizar la mediana --no la media-- para dar una idea cuantitativa de tendencia central. El asunto de los *outliers* es delicado: excluirlos del análisis puede significar una pérdida irreparable en la base de datos, pues en ocasiones son precisamente estos casos extremos o atípicos los que dan particularidad al fenómeno bajo estudio. ¿Qué sería un estudio de municipios del Valle del Cauca sin Cali y sin Palmira? (Ver Ejemplo).

Ejemplo: Los 42 municipios del Valle del Cauca --exceptuando Cali con 2 millones y Palmira con 300 mil-- tienen poblaciones que fluctúan entre 8 mil y 100 mil habitantes. En este caso la media, como indicador de tendencia central para la población **municipal** del Valle se vería fuertemente afectada por los valores de Cali y Palmira, que se comportan como *outliers.* La media daría por tanto una falsa idea. La opción de excluir del conjunto a los dos casos extremos puede significar una pérdida de información preciosa. Sería mejor utilizar la mediana, manteniendo las dos ciudades grandes.

La mediana en general es **poco** utilizada en sociología y debería serlo más, para evitar el abuso de la media que es muy generalizado y puede esconder muchas falacias.