14

Medidas de tendencia central, de dispersión y de concentración

Las *medidas de tendencia central* son números calculados con fórmulas especiales que representan, en forma sumaria, una serie de valores de una variable cuantitativa. Por su lado, las *medidas de desviación* expresan la heterogeneidad u homogeneidad de esos valores. En esos casos, ambas medidas, como variables colectivas que son, caracterizan al colectivo en el cual se dan los correspondientes valores individuales. Así, por ejemplo, si un grupo tiene un promedio de edad de 15.5 años y otro tiene un promedio de 18.6, el primero se caracteriza por su "menor" edad respecto del segundo.

MODA

La *moda* es el valor de una serie que se da con mayor frecuencia entre los miembros de un colectivo. Puede ser utilizado con variables nominales, ya que basta contar los números de sujetos que hay en cada categoría de una variable de este tipo (por ejemplo, el número de hombres y el número de mujeres). Obviamente, es muy fácil de determinar y por ello se le emplea como una primera medida de tendencia central. En la serie siguiente se ve sin problemas que la moda es el número 20:

8, 7, 6, 10, 15, 16, 20, 20, 20, 21, 23

Algunas veces hay más de una moda:

6, 7, 8, 9, 9, 9, 10, 11, 12, 12, 12

En las cifras anteriores hay dos modas: los números 9 y 12. A una distribución como ésa se le denomina *bimodal*.

La facilidad de cálculo de la moda se paga con algunas debilidades:

a) La medida varía considerablemente de una muestra a otra tomada del mismo universo.
b) Puede no dar una buena representación del colectivo del cual proviene. Por ejemplo, las dos distribuciones siguientes, muy distintas entre sí por sus valores componentes, tienen la misma moda (el número 5), con lo cual podría creerse que las dos series de valores son semejantes, cuando, en realidad, hay bastantes diferencias entre ellas:

- 1, 3, 4, 5, 5, 5, 6, 7, 9
- 4, 5, 5, 5, 8, 8, 10, 10, 18, 25

MEDIANA

La *mediana* (md) es el valor que ocupa el lugar central de una distribución ordenada de valores, habitualmente en orden ascendente. Si el número de valores es impar, la mediana es el valor central. Si ese número es par, la mediana es la semisuma de los dos valores centrales. Ejemplos:

- Número impar de valores, ya ordenados: 10, 12, 14, 16, 19; la mediana es 14.
- Número par de valores, ya ordenados: 12, 14, 15, 16, 18, 20. La mediana es la semisuma de los valores centrales, 15 y 16; es decir, 15.5.

Cuando se agrupan los datos en intervalos de clase, se utiliza la fórmula correspondiente que aparece en cualquier texto de estadística.

La mediana es una medida de tendencia central que está especialmente indicada para datos ordinales, como puntajes obtenidos en la medición de actitudes, calificaciones, etc. A diferencia de la media aritmética, que presentamos a continuación, no está influida por valores extremos –muy altos o muy bajos– que se pueden dar en una serie de valores.

MEDIA ARITMÉTICA

La *media aritmética* es una de las medidas de tendencia central más utilizada para caracterizar a un colectivo mediante un solo valor. Ese valor es la suma de los valores de una variable cuantitativa continua, de carácter interval o proporcional, dividida entre el número de valores sumados. Su fórmula para datos no agrupados es la siguiente:

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Por ejemplo, si las cifras siguientes indican el número de horas de cada uno de seis niños que ven televisión al día –2.5, 3, 3, 3.5, 2, 1–, el método aritmético de esa actividad es la suma de las horas dividida entre 6:

$$\bar{x} = \frac{2.5 + 3 + 3 + 3.5 + 2 + 1}{6} = 2.5$$

En algunas oportunidades, de manera incorrecta, se utiliza la media aritmética con datos ordinales (por ejemplo, con calificaciones dadas a los alumnos por el profesor). En tales casos, debe tenerse en cuenta que el valor obtenido es sólo aproximado, por cuanto esos puntajes indican jerarquía entre ellos, y por tanto, los intervalos entre cualquier par de números pueden ser desiguales. Esta observación tiene especial importancia cuando se hacen comparaciones entre medias aritméticas, repetimos, de nivel ordinal, respecto de las cuales el investigador que analiza los datos debe tomar las precauciones del caso.

Como dijimos antes, cuando en una serie de datos hay valores extremos que pueden distorsionar la representatividad de ella (como sería el caso en la serie 3, 6, 8, 21), conviene utilizar la mediana.

VARIANZA Y DESVIACIÓN ESTÁNDAR

Son medidas de dispersión o de variabilidad de los datos de una serie de valores. Indican, como se dijo en la introducción de esta parte, la homogeneidad o heterogeneidad de ellos y, por tanto, la semejanza o diferencia que existe entre los individuos de un colectivo en relación con una cierta variable cuantitativa (la edad, los ingresos, etc.). Las principales de esas medidas son la *varianza*, la *desviación estándar* y el *índice de dispersión*. Las dos primeras deben utilizarse con variables intervales o proporcionales; el índice de dispersión se aplica a variables ordinales y nominales.

La *varianza* es el promedio de las desviaciones elevadas al cuadrado, de cada uno de los valores de una serie respecto de la media aritmética de ella. La *desviación estándar*, a su vez, es la raíz cuadrada de la varianza. Las fórmulas de estas medidas son:

- Varianza:

$$s^2 = \frac{\Sigma(x_1 - x)^2}{n}$$

- Desviación estándar:

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

Analicemos un ejemplo de cálculo. Se ha preguntado a seis niños el número de horas que dedican al estudio fuera de la escuela. Las respuestas son las que se indican a continuación. Calcular la varianza y la desviación estándar de esas horas.

Niño	*Horas de estudio: x*	$(x - \bar{x})$	$(x - x)^2$
1	2	$2 - 2.75 = -0.75$	0.56
2	3	$3 - 2.75 = 0.25$	0.06
3	2.5	$2.5 - 2.75 = -0.25$	0.06
4	3	$3 - 2.75 = 0.25$	0.06
5	2	$2 - 2.75 = -0.75$	0.56
6	4	$4 - 2.75 = 1.25$	1.56
			2.86

Para calcular la varianza y derivar de ella la desviación estándar, la primera tarea consiste en calcular la media aritmética de los valores de las horas:

$$\bar{x} = \frac{2 + 3 + 2.5 + 3 + 2 + 4}{6} = 2.75$$

Luego se hacen las otras operaciones:

Varianza = $2.86 : 6 = 0.48$

Des. est. = $\sqrt{0.48} = 0.69$

Las desviaciones estándar de dos distribuciones de frecuencia no se pueden comparar directamente, pues dependen del tamaño de la media aritmética respectiva. Para hacerlo, hay que expresarlas como porcentajes de esas medias, las cuales reciben el nombre de *coeficientes de variación*. Su fórmula de cálculo es la siguiente:

$$v = \frac{s}{\bar{x}} \times 100$$

Por ejemplo, supongamos que, en un cierto grupo, el promedio de las edades es de 26 años, con una desviación estándar de 3. En otro, el promedio es de 38 años, con una desviación estándar de 5. Los coeficientes de variación son, respectivamente, de $3 : 26 \times 100 = 11.5$ y $5 : 38 \times 100 = 13.2$. Si se hubiesen comparado directamente las desviaciones estándar, se podría haber dicho que la dispersión era mucho mayor en el segundo grupo que en el primero, pues la desviación del caso era de 5 contra 3 (1.7 veces más). En cambio, los coeficientes de variación muestran una diferencia menor (1.1 veces más).

MEDIDAS DE CONCENTRACIÓN DE UNA VARIABLE

Para determinar la concentración que puede tener una variable cuantitativa en un cierto colectivo, se utilizan dos medidas principales: una de ellas es el *índice de Gini* y la otra es la diferencia de la variable entre *quintiles* extremos de la distribución.

Índice de Gini

Supongamos que deseamos averiguar cuál es el grado de concentración de la educación en una población de personas de la cual conocemos los siguientes datos:

Tipo de ocupación	*Porcentaje respecto del total*	*Porcentaje de personas con educación universitaria*
Obreros rurales	17.0	2.5
Obreros urbanos	38.5	4.8
Agricultores	22.8	5.1
Empleados	11.8	20.9
Empresarios	7.2	26.9
Profesionales y técnicos	2.7	30.8
	100.0	100.0

Para el cálculo de la concentración se usa el índice de Gini, cuya fórmula de cálculo es:

$$\text{Gini} = 1 - \Sigma(p_i + p_{i-1}) \times (q_i + q_{i+1})$$

en la cual:

p_i = proporciones de personas acumuladas en cada grupo (tipo de ocupación, en el ejemplo).
q_i = proporciones acumuladas de la participación de cada grupo en la variable del caso (educación, en el ejemplo).

El coeficiente varía de 0 a 1 (1 es la concentración máxima).

Veamos un ejemplo de cálculo. (Aquí, los porcentajes se expresan en proporciones.)

p_i	q_i	p_{i+1}	$q_i + q_{i+1}$	$(p_i - p_{i+1})(q_i + q_{i+1})$
0.170	0.025	–	–	–
0.555	0.073	0.385	0.098	0.04
0.783	0.124	0.228	0.197	0.04
0.901	0.333	0.118	0.457	0.05
0.973	0.602	0.072	0.935	0.06
1.000	1.000	0.027	1.602	0.04
				0.23

El índice de Gini es de $1 - 0.23 = 0.77$, valor que indicaría alta concentración o desigualdad de la educación superior en el grupo estudiado (ficticio).

El coeficiente de Gini es la expresión matemática de la *curva de Lorenz*, que resulta de representar en un eje cartesiano los porcentajes de población y de los de la variable cuya concentración se desea conocer. Ésta corresponde al área que queda entre la diagonal del diagrama –diagonal que expresa una distribución perfecta– y la curva de Lorenz, que indica la distribución real observada en la población del caso. En la figura 14.1 (tomada de Sierra Bravo, p. 481) se puede apreciar una situación como esta última.

Frecuencias acumuladas

Porcentaje acumulado pob.	*Porcentaje acumulado univ.*
17.0	2.5
55.5	7.3
78.3	12.4
90.1	33.3
97.3	60.2
100.0	100.0

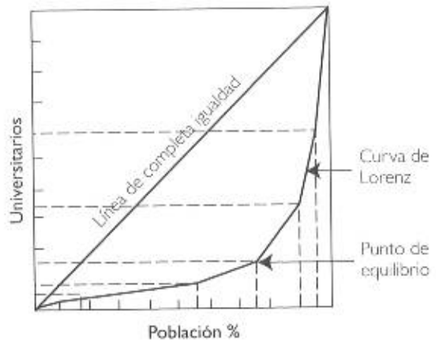

Fig. 14.1. Curva de Lorenz.

Quintiles

Es otro procedimiento para determinar si existe o no concentración de una variable en un colectivo. Consiste en dividir en quintiles el total de la población (grupos de 20 % de ella); con los valores de la variable que determinan esa división se calculan los totales de la variable que queda en cada grupo. El nivel de concentración se comprueba comparando los valores que quedan en los quintiles extremos. Como ejemplo, supongamos la siguiente situación, en relación con la escolaridad de una comunidad, con los valores de la variable que se indican en cada quintil. El total de la segunda columna resulta de multiplicar el número de personas (que en cada quintil corresponde a 20 % del total) por los valores de la variable del quintil correspondiente.

Quintiles (en años)	*Total en años de escolaridad en cada quintil*	*Porcentaje respecto del total de año de escolaridad*
Primero 1-2 3-4 . . .	96 200	12.9
Quinto 15-16 17-18	224 960	30.3
		Total 100.0 (742 960)

Las cifras utilizadas en el ejemplo muestran que mientras 20 % con menor escolaridad recibe 12.9 % del total de la escolaridad que existe en toda la comunidad, 20 % del quintil más alto recibe más del doble de esa escolaridad: 30.3 %.

Este procedimiento puede utilizarse dividiendo la distribución en quintiles de otra variable para calcular en cada uno de ellos el total de la variable cuya concentración se desea conocer. Así se puede ver en el siguiente ejemplo, que muestra la concentración de la educación superior en cada quintil de ingreso y los cambios ocurridos en dos periodos (Ministerio de Planificación de Chile, *Programas Sociales: Su impacto en los hogares chilenos*, 1990, pp. 62-63).

Distribución comparativa 1987-1990 de la matrícula de educación superior, según quintil de ingreso per cápita

Quintil de ingreso per cápita	1987		1990		*Variación de porcentaje* $= \frac{n_2}{n_1}$
	n_1	*Porcentaje*	n_2	*Porcentaje*	
1	12 243	5.2	22 709	9.1	85.5
2	19 495	8.3	29 605	11.9	51.9
3	40 993	17.4	41 869	16.8	2.1
4	63 766	27.1	59 883	24.1	−6.1
5	99 154	42.1	94 765	38.1	−4.4
Total	235 651	100.0	248 831	100.0	5.6

FUENTE: ODEPLAN, *Encuesta CASEN 1987*, Departamento de Planificación y Estudios Sociales, MIDEPLAN, *Encuesta CASEN 1990*.

De acuerdo con la información que proporciona la encuesta CASEN, la matrícula total de educación superior ha aumentado de 235 651 alumnos en 1987 a 248 831 en 1990, lo que representa un incremento promedio de 5.6 %. Si se analiza la distribución de la matrícula según el nivel de ingreso per cápita del hogar, se puede apreciar que en ambos casos la distribución es bastante heterogénea: las diferencias entre los quintiles extremos alcanzan 36.9 puntos porcentuales en 1987 y 29.0 puntos en 1990.

Entre 1987 y 1990 se constatan cambios significativos en la distribución de la matrícula de educación superior según quintil de ingreso. La matrícula del primer quintil aumenta de 5.6 % en 1987 a 9.1 % en 1990; en el segundo, de 8.3 % a 11.9 %. En el tercero, cuarto y quinto quintiles, que corresponden a los más altos ingresos, se produce una disminución que fluctúa entre menos de 1 % y 4 %. Este aumento de matrícula, que se concentra fundamentalmente en los quintiles de menores ingresos, representa un incremento de 85.5 % en la matrícula del primer quintil y de 51.9 % en el segundo. En los quintiles de mayores ingresos, en cambio, representa una disminución en la matrícula de 6.1 % y 4.4 %, respectivamente.

BIBLIOGRAFÍA SELECCIONADA PARA LA QUINTA PARTE

Blalock, Hubert, *Estadística social*, Fondo de Cultura Económica, México, 1996, segunda parte.

Ferrán, Magdalena, *SPSS para Windows. Programación y análisis estadístico*, McGraw-Hill, Madrid, 1996, parte II: 4. Estadística descriptiva.

Loether, Herman y Donald G. MacTavish, *Descriptive Statistics for Sociologists*, Allyn and Bacon, Boston, 1974.

Salkind, Neil, *Métodos de investigación*, Prentice Hall, México, 1999, cap. 7.

EJERCICIO 5 (AUTOEVALUACIÓN)

1. ¿Cuáles son las principales funciones que cumplen las razones, proporciones, porcentajes y tasas en el análisis de datos?
2. ¿En qué casos es conveniente utilizar una proporción en lugar de una razón, al tratar la distribución de casos entre las categorías de una variable cualitativa?
3. ¿Qué dice la regla de Zeisel?
4. ¿Cuáles son los métodos más utilizados para determinar una tendencia en una serie de tiempo?
5. Calcule los efectos de las variables "antigüedad en el cargo" y "autoestima profesional" en el siguiente cuadro que corresponde a 208 empleados en una empresa.

	Autoestima		
Antigüedad (años)	*Alta*	*Mediana*	*Baja*
15-20	15	20	25
19-14	32	30	15
13 y menos	28	25	18

6. En la población del ejercicio anterior, se tiene la siguiente información sobre porcentajes de educación superior en cada grupo de antigüedad:

15-20	20%
19-14	35%
13 y menos	45%

7. Calcule en esa población, mediante el índice de Gini, el grado de concentración de la educación superior. Comente el resultado.

Parte VI

Medidas de asociación y de correlación

Como hemos dicho en el capítulo anterior, el cálculo de relaciones entre dos variables forma parte del análisis descriptivo. La relación entre más de dos variables cae convencionalmente en el análisis explicativo, en la forma que lo veremos al desarrollar el análisis multivariable. En esta parte abordaremos las relaciones que pueden darse entre variables nominales, ordinales, intervales y de razón, con sus particulares características y formas de cálculo. En la bibliografía pertinente suele darse el nombre de *asociación* a la relación entre dos variables nominales o entre dos variables ordinales. En cambio, se usa el término *correlación* para designar la relación entre dos variables intervales o de razón.

En la relación entre variables ordinales, intervales y de razón, una de ellas es tomada como variable independiente, y la otra, como variable dependiente. En tal relación se consideran cuatro aspectos importantes:

a) La *existencia o no de relación*. Existe relación cuando, al cambiar los valores o categorías de una variable, también, concomitantemente, cambian los valores o categorías de la otra variable.
b) *Grado de relación entre las variables*. Indica la magnitud de la relación. En la mayoría de los casos, los coeficientes de la relación varían entre los valores −1 y +1.
c) *Dirección de la relación*. La relación es positiva cuando, al aumentar el valor de una variable (o de una categoría de ella), el valor (o categoría) de la otra también aumenta. En cambio, si al aumentar el valor de una de las variables el valor (o categoría) de la otra disminuye, la relación es negativa.
d) *Naturaleza de la relación*. La relación es lineal cuando, al aumentar o disminuir de manera uniforme los valores de una de las variables, los valores de la otra también aumentan o disminuyen en igual proporción. La relación es *curvilínea* cuando, al aumentar o disminuir los valores de una variable, la otra lo hace en sentido contrario.

Las asociaciones se calculan en las denominadas *tablas de contingencia*, es decir, en tablas que resultan del cruce de dos variables, cada una con sus correspondientes categorías, las cuales definen casillas en el cuadro del caso. Así, para dar un ejemplo sencillo, si tenemos la variable "sexo", con sus categorías "hombre" y "mujer", y la variable "religiosidad", con las categorías "alta" y "baja", su cruzamiento define las categorías bivariadas: "hombre con alta religiosidad", "mujer con alta religiosidad", "hombre con baja religiosidad" y "mujer con baja religiosidad". En general, la asociación se busca en cuadros con r filas y c columnas, correspondientes a dos variables x y con categorías r y c, respectivamente.

El cuadro –resumen que hemos tomado de Magdalena Ferrán (1996, pp. XXX a XXXII)– presenta las principales medidas de asociación que pueden darse entre dos variables de naturaleza nominal y ordinal.

Medidas de asociación para tablas de contingencia

Medida de asociación	*Tabla*	*Escala de medida*	*Observaciones*
▪ Phi	2×2	nominales	▪ Son medidas basadas en el estadístico chi cuadrado.
▪ Coeficiente de contingencia	$r \times c$	nominales	▪ Toman valores comprendidos entre 0 y 1, que indican un mínimo y máximo grado de asociación, respectivamente.
▪ V de Cramer	$r \times c$	nominales	▪ Phi presenta el inconveniente de que puede alcanzar valores superiores en tablas de $r \times c$; el coeficiente de contingencia depende de una cota superior, y la V de Cramer tiende a subestimar la asociación. Además, pueden tomar el mismo valor en muestras con tamaños muy diferentes.
			▪ Son útiles para comparar grados de asociación entre pares de variables observadas sobre un mismo conjunto de individuos.

Medidas de asociación para tablas de contingencia (*Continuación.*)

Medida de asociación	*Tabla*	*Escala de medida*	*Observaciones*
▪ Riesgo relativo	2 × 2	nominales	▪ Toma valores positivos. Si las variables son independientes, su valor será próximo a 1. ▪ Compara los dos grupos establecidos por los valores de una de las variables, en función de la frecuencia con que presentan cada uno de los valores de la otra. ▪ Admite la posibilidad de distinguir entre grupos de control experimental.
▪ Lambda	*r* × *c*	nominales	▪ Toman valores comprendidos entre 0 y 1, que indican mínimo y máximo grados de asociación, respectivamente. ▪ Disponen de versión asimétrica. ▪ Lambda es fácil de interpretar en función de la proporción en que se reduce el error en la predicción del valor de una variable a partir de los valores de la otra; sin embargo, puede tomar el mínimo valor de tablas con asociación. ▪ El coeficiente de incertidumbre únicamente toma el valor cero en tablas con no asociación; sin embargo, su valor es más difícil de interpretar que el de lambda.
▪ Coeficiente de incertidumbre	*r* × *c*	nominales	
▪ Kappa	*r* × *r*	nominales	▪ Los posibles valores de las dos variables son los mismos.

			▪ Toma valores comprendidos entre −1 y 1, que indican, respectivamente, mínimo y máximo grados de acuerdo entre los valores de las dos variables.
▪ Gamma	*r* × *c*	ordinales	▪ Toma valores comprendidos entre −1 y 1, que indican máximo grado de asociación negativa y positiva, respectivamente. ▪ Gamma es fácil de interpretar, pero puede alcanzar valores extremos en tablas en las que la asociación no es total.
▪ Tau *b* de Kendall	*r* × *c*	ordinales	▪ Únicamente alcanza valores extremos en tablas con asociación total; sin embargo, si *r* es distinto de *c*, no puede alcanzarlo.
▪ Tau *c* de Kendall	*r* × *c*	ordinales	▪ Puede alcanzar valores extremos aun en el caso de que *r* sea distinto de *c*; sin embargo, tiende a subestimar la asociación.
▪ *D* de Somers	*r* × *c*	ordinales	▪ Dispone de distribución asimétrica; sin embargo, puede alcanzar valores extremos en tablas en las que la asociación no es total.