García Ferrando

LOETHER, Herman J., y Donald G. MCTAVISH: *Descriptive Statistics for Sociologists*, Boston, Allyn and Bacon Inc., 1974.
MARTÍN MARTÍNEZ, José Luis: «Ensayo de tipificación de los sin opinión», *Revista Española de Investigaciones Sociológicas*, 16, 1981, págs. 9-37.
MOSIMANN, Thomas F.: «Mathematical Statistics and Real Statistics», *IASI, Estadística*, junio 1957, págs. 390-394.
STEVENS, S. S.: «Mathematics, Measurement and Psychophysics», en S. STEVENS (ed.): *Handbook of Experimental Psychology*, New York, Wiley, 1951, págs. 1-30.
THURSTONE, L. L.: «Attitudes can be measured», *American Sociological Review*, vol. 33, 1928, págs. 529-554.

Capítulo 2

ESTADISTICA DESCRIPTIVA UNIVARIABLE: LA LOGICA DEL ANALISIS COMPARATIVO

2.1. IMPORTANCIA DEL ANÁLISIS COMPARATIVO EN LA SOCIOLOGÍA

Con frecuencia, escuchamos y vemos mensajes publicitarios que reclaman para sus productos aspectos cuantitativos con los que se pretende atraer la atención del público. Así, una marca de cigarrillos anuncia que su tabaco contiene un tanto por ciento determinado menos de nicotina. Aunque no se nos dice claramente, se sobreentiende que con el mensaje se pretende afirmar que tal marca de cigarrillos es mejor que otras que se encuentran en el mercado. Sin embargo, y desde una perspectiva estrictamente lógica, no se puede inferir que una marca de cigarrillos sea mejor que otra —desde el punto de vista de su contenido en nicotina— a partir del porcentaje que se anuncia, porque está ausente todo elemento comparativo y no se puede interpretar debidamente dicho porcentaje.

Así, con el mensaje publicitario que se transmite, ¿qué se pretende afirmar? ¿Que los cigarrillos tienen ahora menos nicotina que la que contenían hace un año? ¿Que dicho contenido está por debajo de la media de otras marcas competidoras? ¿O que se encuentra por debajo del contenido en nicotina que se puede considerar perjudicial para la salud? El problema, pues, que surge con tal mensaje publicitario es que contiene implícitamente una comparación, pero sólo ofrece parte de la información. Sin un referente o una medida estándar, no es posible extraer ninguna conclusión válida sobre el porcentaje nicotínico de menos que contienen los cigarrillos.

Veamos otro ejemplo de falta de elementos comparativos en los que justificar el resultado. Un semanario español de gran tirada encabezaba el informe de una encuesta sobre actitudes sexuales de la población española con el siguiente título: «Jóvenes españoles: más progres que nadie» (*Cambio 16*, núm. 311, 1977). Esta afirmación se basaba en el siguiente resultado: el 72 por 100 de los jóvenes españoles de ambos sexos considera como algo correcto el tener relaciones sexuales sin estar casados; un 27 por 100 afirma que no las tendría personalmente, y otro

26 por 100 lo juzga como algo incorrecto. No obstante, estudios similares realizados en otros países de la Europa occidental ponen de manifiesto que los porcentajes de jóvenes que no tendrían relaciones sexuales sin estar casados son sensiblemente menores que en el estudio español. Concretamente, el 12 por 100 en Finlandia, el 10 por 100 en Francia, el 21 por 100 en Grecia, el 26 por 100 en Italia y el 19 por 100 en Gran Bretaña. La comparación de tales porcentajes parece invalidar, pues, la afirmación de que los jóvenes españoles son «más progres que nadie».

Nuestra vida cotidiana, tanto la de los científicos sociales como la de cualquier otro ciudadano, está repleta de informaciones que contienen datos que, de algún modo, reclaman la realización de comparaciones. La tasa bruta de natalidad por 1.000 habitantes en España, en el período 1971-73, es de 19,4. ¿Se trata de una tasa alta o baja? ¿Es mayor, igual o menor que las correspondientes tasas de otros países europeos? Para responder a estas y otras preguntas se requiere una información adicional con el objeto de realizar una comparación válida.

Toda investigación comporta problemas de comparaciones, al tratar de alcanzar conclusiones relevantes, tal como sugieren los siguientes ejemplos:

> «El Censo de 1950 nos indica que las mujeres casadas o que han estado casadas han tenido, como media, 3,1 hijos. El mismo cálculo, veinte años después, nos señala un promeio de 2,8 hijos. Es evidente, por tanto, el descenso general de la fecundidad a lo largo del período 1950-1970» (Amando de Miguel, 1977, pág. 52).
>
> «El número total de becarios del Patronato de Igualdad de Oportunidades del Ministerio de Educación ha pasado de 34.246, en 1961, a 268.000 en 1971. Esto quiere decir que ha aumentado en un 688 por 100, mientras las asignaciones económicas totales han crecido un 433 por 100. Lo cual indica que ha bajado la cuantía de las asignaciones individuales, aparte de la depreciación de su valor adquisitivo, en diez años» (FOESSA, 1975, 242).
>
> «De febrero de 1976 a febrero-marzo de 1978, el juicio sobre la situación económica del país se va tiñendo de colores más sombríos, como no podría esperarse otra cosa. Si en el 76 era un 64 por 100 el que evaluaba la situación económica como mala o muy mala, en el 78 ese porcentaje llega hasta el 80 por 100» (Andrés Orizo, 1979, pág. 63).

A través de los ejemplos anteriores se observa que las conclusiones se extraen a partir de las comparaciones realizadas sobre los resultados obtenidos en fechas diferentes, y es que la sociología hace continuamente uso de las comparaciones para avanzar el pensamiento sociológico. Ya Emile Durkheim, en *Las reglas del método sociológico*, publicado originalmente en 1895, afirmaba que «la sociología comparada no es una rama especial de la sociología; es la sociología misma, en tanto que cesa

de ser meramente descriptiva y aspira a explicar los hechos». Para el sociólogo francés, cuando la producción experimental de los hechos no es posible —que es lo más usual en la investigación sociológica—, el método de investigación que hay que seguir es el comparado (Durkheim, 1972, págs. 163 y sigs.).

De este modo queda puesto de manifiesto la importancia del análisis comparativo en la investigación sociológica. Llegados a este punto, cabe hacerse dos preguntas en relación al análisis comparativo: *¿Qué* debe compararse? *¿Cómo* debe hacerse la comparación?

2.2. Tipos de comparación

La respuesta a la pregunta de qué cosas o fenómenos deben compararse depende, básicamente, de lo que se esté estudiando. Se trata de un tema estrictamente teórico. Si el problema de investigación está claramente formulado y conceptualizado, será más fácil saber qué datos es preciso reunir para realizar la comparación. Sin una conceptualización clara del problema será muy difícil elegir de entre las muchas alternativas que nos ofrece un entorno social cada vez más repleto de informaciones de todo tipo. Por ello se hace preciso formular y medir con todo cuidado las variables e identificar con toda claridad los objetos o fenómenos que se van a medir para que se puedan contrastar consistentemente grupos comparables. A lo largo de este libro insistiremos, siempre que sea oportuno, en que el análisis estadístico sólo puede ser relevante y fructífero una vez se hallan resuelto, al menos suficientemente, los problemas de teoría, conceptualización, medición y diseño que toda investigación comporta. En caso contrario, el análisis estadístico sólo servirá para dar una mera apariencia de seriedad y de rigor a unos resultados que probablemente serán inciertos, toda vez que no se hallan resueltos previamente los problemas teóricos y metodológicos a los que nos hemos referido.

De una forma general, tres son los tipos de comparación que se pueden realizar:

1. Comparaciones entre grupos, bien sea dentro del mismo estudio o entre estudios diferentes.
2. Comparaciones entre un grupo y un caso individual de dicho grupo.
3. Comparaciones entre los resultados de un estudio y unos resultados estandarizados que o bien han sido establecidos a partir de investigaciones previas o provienen de un modelo teórico formulado por el investigador.

2.2.1. *Comparaciones entre grupos*

El modelo de comparación ideal, desde un punto de vista científico, es el realizado entre un grupo experimental al que se le ha sometido a un tratamiento conocido, como podría ser un grupo de alumnos al que se le enseña un programa educativo especial, y un grupo de control que no ha sido sometido a dicho tratamiento —en nuestro ejemplo, sería un grupo de alumnos al que se le continúa enseñando un programa tradicional.

Este tipo de comparación entre un grupo experimental y un grupo de control está relacionado con el modelo de un experimento controlado, que constituye el diseño científico ideal. Este diseño consiste sencillamente en la comparación realizada entre un grupo experimental y un grupo de control en dos momentos en el tiempo, esto es, antes y después de someter el primer grupo al tratamiento especial o experimento. En un breve pero sustancial artículo, el sociólogo americano Samuel A. Stouffer (1950) destaca la escasa frecuencia con que los sociólogos emplean dicho modelo de diseño en sus investigaciones, utilizando, por el contrario, diseños que sólo incluyen dos observaciones, y a veces sólo una, en vez de las cuatro que se requieren en el modelo experimental. Naturalmente, los resultados científicos que se pueden obtener de diseños tan limitados no pueden ser muy esperanzadores.

Los grupos que se comparan pueden estar constituidos por individuos o por cosas u objetos no personales, tales como grupos de organizaciones o instituciones sociales. Los grupos que se comparan suelen venir caracterizados por una serie de puntuaciones sobre medidas de dimensiones o aspectos definidores del grupo. En tal caso, lo primero que hay que hacer es resumir dichas puntuaciones por medio del estadístico que se considera más oportuno, comparándose de este modo los estadísticos resúmenes de cada grupo.

2.2.2. *Comparación entre un grupo y un individuo*

Otro tipo de comparación es la que se realiza entre un grupo y un individuo —o caso individual— que forma parte del grupo. Así, podemos comparar la tasa de delincuencia en una ciudad con la tasa media correspondiente a la sociedad en general, o bien comparar la conflictividad laboral de una empresa determinada con la que muestra el sector productivo en el que se inscribe la empresa. En el caso de personas, se puede comparar los resultados escolares de un estudiante con los correspondientes a la media de la clase a la que asiste dicho estudiante. Lo importante en todos los casos mencionados consiste en delimitar y definir las características del grupo que se compara con las correspondientes al individuo.

2.2.3. *Comparación entre el resultado de un estudio y un resultado estándar*

Por último, se pueden comparar los resultados obtenidos en un estudio determinado con unos resultados estándar. Así, se pueden contrastar determinadas características demográficas de un grupo social objeto de nuestro estudio con las correspondientes tasas que ofrecen los resultados del Censo General de Población. Otras veces, el estándar es simplemente un estudio anterior que sirve de referente a una nueva investigación, como podría ser el caso de un antropólogo que estudia una comunidad rural que ya ha sido estudiada anteriormente por otro colega.

Conviene destacar aquí las comparaciones que pueden realizarse a partir de las teorías conocidas. De hecho, las teorías son una fuente sugerente de comparaciones estándar. Sabemos, por ejemplo, que la teoría de la transición demográfica de las sociedades que pasan del estado preindustrial al industrial predice un cambio en las tasas de natalidad y de mortalidad, de forma que los valores altos de tales tasas se reducen significativamente. Pues bien, podemos comparar la evolución en el tiempo de las tasas demográficas correspondientes a una sociedad concreta, para observar los cambios que se están produciendo en ella desde el punto de vista de la teoría de la transición demográfica, y cómo se está alterando la pirámide de población y la tasa de crecimiento demográfico de dicha sociedad en el período considerado.

También sabemos, por la teoría de la estratificación social, que los grupos sociales en los que predominan los individuos con una elevada inconsistencia de *status* son potencialmente más conflictivos e inestables que los grupos sociales en los que predominan los individuos con unos componentes de *status* más equilibrados. Pues bien, podemos estudiar un grupo de población determinado desde el punto de vista de su inconsistencia de *status* y el grado de conflictividad e inestabilidad que manifiesta. En resumen, pues, la teoría sociológica está repleta de resultados que pueden servirnos para contrastar los hallazgos de nuevas investigaciones. Tales comparaciones servirán, además, para contrastar, en el sentido de modificar, rechazar o modificar, la teoría que sirve como comparación estándar.

2.3. Operaciones básicas de comparación

Los procedimientos existentes para realizar las operaciones de comparación son muy variados. De hecho, el campo de la estadística descriptiva, que es el más amplio y comúnmente utilizado por los sociólogos —en relación con el campo de la estadística inferencial—, tiene como uno de sus temas recurrentes la realización de comparaciones significativas entre agrupaciones de datos cuantitativos.

La realización de tales comparaciones en el campo de la estadística descriptiva incluye dos operaciones fundamentales. La primera de ellas se refiere a la organización y ordenación de los datos o medidas obtenidos en algún tipo de distribución, mientras que la segunda de dichas operaciones se refiere al tratamiento aritmético de dichos datos, bien sea por medio de la resta o sustracción o bien por medio de la división. Tal como destacan Loether y McTavish (1974, 43), la idea de la división, esto es, la creación de una relación entre un número (numerador) y otro número (denominador), es uno de los temas organizadores básicos tanto de la estadística descriptiva como de la estadística inferencial. Desde el punto de vista de la relevancia que tiene para la investigación sociológica la creación de tales relaciones, el problema consiste en saber qué es lo que hay que dividir entre qué, y la respuesta, normalmente, vendrá dada por el esquema teórico en el que se enmarque la investigación.

A continuación vamos a presentar un breve panorama de las operaciones básicas de comparación, utilizando para ello ejemplos prácticos de carácter sociológico.

2.3.1. *La organización de los datos*

Una lista de datos que no esté organizada según un criterio determinado suele ser de poca utilidad para el investigador interesado en realizar algún tipo de comparación. Una vez se hayan obtenido los datos que estimamos relevantes para realizar el análisis deseado es conveniente ordenarlos según algún criterio, bien sea de mayor a menor o de otra forma, con el fin de que se pueda obtener el máximo de información posible de los datos. La ordenación permitirá observar con mayor facilidad la distribución de los datos y el lugar dónde termina un grupo y comienza otro en relación a otros grupos.

Supongamos que estamos estudiando la población extranjera de origen europeo residente en España. La primera información que necesitamos reunir es la referente al número y origen de esta población extranjera. El *Anuario Estadístico* del Instituto Nacional de Estadística ofrece los datos que se recogen en la tabla 1 sobre la nacionalidad de origen y el número de extranjeros que han residido en España en 1979.

Los países se presentan ordenados en el *Anuario* por orden alfabético y así los hemos transcrito en la tabla 1. Una ordenación de este tipo puede que no resulte la más interesante para ofrecer de una forma relevante la información. Cabría pensar en realizar otra ordenación de los países según el número de personas que tienen residiendo en España. Así, tendríamos encabezando la lista a Portugal, con 21.801 personas, seguido de Alemania, con 18.144, y en el otro extremo estarían los países con menor número de residentes, que son la URSS, con 22, y Rumanía, con 26.

TABLA 1

Extranjeros europeos residentes en España, según nacionalidades (1979)

Nacionalidad	*Número*	*Nacionalidad*	*Número*
Alemania	18.144	Italia	9.192
Austria	1.145	Noruega	749
Bélgica	3.764	Países Bajos	4.784
Dinamarca	2.009	Polonia	88
Finlandia	892	Portugal	21.801
Francia	14.891	Rumanía	26
Gran Bretaña	17.330	Suecia	3.229
Grecia	425	Suiza	3.576
Hungría	36	U.R.S.S.	22
Irlanda	376	Yugoslavia	33

FUENTE: *Anuario Estadístico de España*, Madrid, I.N.E., 1979.

Ahora bien, teniendo en cuenta que el número de extranjeros que residen en un país dependerá, entre otros factores, del tipo de relaciones que guarden los países entre sí, es decir, de su proximidad política, cultural y económica, aparte de su proximidad geográfica, cabe realizar una ordenación de los países en grupos regionales, tal como se presenta en la tabla 2.

Con la nueva agrupación se obtiene, a primera vista, una ordenación más significativa de los datos. Así, se observa que el grupo más amplio de extranjeros, con 70.889 personas, proviene de los países de la Europa occidental, con los que España mantiene unos estrechos contactos de todo tipo. Este elevado grupo de europeos occidentales contrasta con el pequeño grupo de extranjeros que provienen de los países europeos socialistas, sólo 205, con los que España mantiene unas relaciones mucho más escasas y distanciadas. El resto de los países lo hemos distribuido entre países mediterráneos, que incluye a Grecia e Italia, con 9.617 residentes en España —la mayoría de ellos italianos—, y Portugal, que, como país vecino, lo hemos mantenido en una categoría aparte, además de tener el máximo número de residentes extranjeros en España.

La agrupación realizada, pues, ha permitido realizar una comparación con la que se pueden analizar de forma más relevante los datos originales. El marco teórico en el que se inscriba el análisis cuantitativo debe ser, en toda investigación empírica, el criterio básico que se ha de seguir para agrupar los datos y poder realizar una comparación significativa.

TABLA 2

Extranjeros europeos residentes en España, según áreas regionales (1979)

Resto Europa Occidental		*Península Ibérica*		*Países mediterráneos*		*Países socialistas*	
Alemania	18.144	Portugal	21.801	Grecia	425	Polonia	88
Inglaterra	17.330			Italia	9.192	Hungría	36
Francia	14.891	TOTAL	21.801	TOTAL	9.617	Yugoslavia	33
Holanda	4.784					Rumanía	26
Bélgica	3.764					URSS	22
Suiza	3.576					TOTAL	205
Suecia	3.229						
Dinamarca	2.009						
Austria	1.145						
Finlandia	892						
Noruega	749						
Irlanda	376						
TOTAL	70.889						

FUENTE: I.N.E., *op. cit.*, elaboración propia.

2.3.2. *Distribuciones*

Con el fin de obtener una organización más resumida y operativa de los datos, se utilizan tres tipos de distribuciones: *a)* la distribución de frecuencias; *b)* la distribución porcentual; *c)* la distribución acumulada.

2.3.2.1. *Distribución de frecuencias*

Cuando se está manejando un número amplio de datos, resulta conveniente distribuirlos en *clases* o *categorías* y determinar el número de casos que pertenece a cada clase. Este número se denomina *frecuencia de clase*, y se simboliza por medio de la letra f o f_i, en donde i se refiere a la clase i de la variable ordenada. El número total de casos es igual, por tanto, a la suma de la columna de las frecuencias, y se simboliza por la letra N, o bien como Σf_i, en donde Σ, que es la letra griega sigma, simboliza la suma de todas las frecuencias de clase.

El número de clases o categorías que se seleccionan vendrá determinado por las necesidades de la investigación. Supongamos que tenemos un grupo de 120 individuos adultos mayores de dieciocho años y menores de setenta y cinco años y queremos distribuirlos según su edad. Una distribución útil puede ser la siguiente:

Edad (años)	f_i
De 18 a 20	10
De 21 a 25	14
De 26 a 35	23
De 36 a 45	20
De 46 a 60	29
De 61 a 75	24
	$N=120$

La primera clase o categoría de edad es «de 18 a 20 años», y a ella pertenecen 10 individuos, es decir, que la frecuencia de esta clase o categoría es 10.

Los datos, tal como han sido ordenados y resumidos en la distribución de frecuencia anterior, se suelen denominar *datos agrupados*. Aunque con el proceso de agrupamiento se pierde algo de la información que contienen los datos originales —por ejemplo, en la categoría «18 a 20» no sabemos cuántos individuos tienen dieciocho, diecinueve o veinte años—, sin embargo, ofrece la gran ventaja de presentar todos los datos de una forma sencilla en un pequeño cuadro, lo que facilita, evidentemente, su estudio.

Continuando con la terminología que se utiliza en la distribución de frecuencia, denominamos *intervalo de clase o categoría* al símbolo que define una clase o categoría; por ejemplo, la clase «de 21 a 25» la simbolizamos como 21-25. Los números extremos de cada clase o categoría, en este caso 21 y 25, se denominan *límites de clase*, siendo el mayor de ellos el *límite superior* y el menor el *límite inferior*. Los términos clase o categoría e intervalo de clase o categoría, que, al menos teóricamente, no tienen límite superior e inferior, se conocen como *intervalo de clase o categoría abierto*. Así, podemos escribir la anterior distribución de frecuencias dejando abierta la categoría «menos de 21» y «más de 60»:

Edad (años)	f_i
Menos de 21	10
De 21 a 25	14
De 26 a 35	23
De 36 a 45	20
De 46 a 60	29
Más de 60	24
	N = 120

Si las edades se registran con una aproximación de meses, el intervalo de la categoría 21-25 incluye, teóricamente, todos los individuos con edades que van desde 20,5 a 25,5 años. Estos números se conocen con la denominación de *límites reales o verdaderos de clase o categoría*, siendo el menor de ellos el *límite real inferior* y el mayor de ellos el *límite real superior*. En la práctica, los límites reales de clase o categoría se obtienen sumando al límite superior de un intervalo de clase o categoría el límite inferior del intervalo contiguo superior y dividiendo a continuación por dos.

Los límites reales se pueden utilizar igualmente para simbolizar las clases o categorías. Así, las diversas categorías del ejemplo anterior podrían indicarse por 17,5-20,5, 20,5-25,5, 25,5-35,5, etc. No obstante, esto se hace raramente, ya que con dicha simbolización se introduce un elemento perturbador por su ambigüedad, ya que los límites reales no coincidirán siempre con las observaciones reales. Por ejemplo, para la edad 25,5 no se puede saber si pertenece al intervalo de la categoría 20,5-25,5 o a la 25,5-35,5. Por esta razón resulta aconsejable utilizar intervalos cuyos límites sean mutuamente excluyentes para las diversas clases o categorías.

El tamaño o amplitud de la clase o categoría es la diferencia entre los límites reales que forman cada clase o categoría, y se conoce como *amplitud, tamaño o longitud* de clase o categoría, según los autores. El tamaño de cada categoría puede ser idéntico o diferente. En la distribución de frecuencia de edades utilizada anteriormente aparecen cuatro tamaños de categoría diferentes, una de tres, otra de cinco, dos de diez y otras dos de quince años, como se observa a continuación: 20,5 — 17,5 = 3; 25,5 — 20,5 = 5; 35,5 — 25,5 = 10; 45,5 — 35,5 = 10; 60,5 — 45,5 = 15; 75,5 — 60,5 = 15.

El punto medio del intervalo de clase o categoría se obtiene sumando los límites inferior y superior de la clase o categoría y dividiendo a continuación por dos. También se denomina *punto medio de la clase o categoría*, y se simboliza por X_i. Así, el punto medio del intervalo 21-25 es $(21+25)/2=23$. En los cálculos estadísticos ulteriores, las observaciones pertenecientes a un intervalo de categoría dado se suponen que son coincidentes con el punto medio de la categoría. Así, todas las edades del intervalo de la categoría 21-25 se considerarán como de edad de veintitrés años. La anterior distribución de frecuencias según categorías de edad se puede escribir del siguiente modo, incluyendo límites reales y puntos medios:

Edad (años)	f_i	Punto medio X_i	Límites reales
De 18 a 20	10	19	17,5 a 20,5
De 21 a 25	14	23	20,5 a 25,5
De 26 a 35	23	30,5	25,5 a 35,5
De 36 a 45	20	40,5	35,5 a 45,5
De 46 a 60	29	53	45,5 a 60,5
De 61 a 75	24	68	60,5 a 75,5
	N = 120		

Tal como se ha señalado anteriormente, el agrupamiento de datos no sólo reporta ventajas, tales como la de resumir y permitir un manejo más fácil de la información, como también presenta algún inconveniente, siendo el principal lo que se denomina *error de agrupamiento*. Con este término nos referimos a las alteraciones que se producen al realizar determinados agrupamientos, lo que conduce a la variación de N. Veamos a través de un ejemplo la aparición de este tipo de error:

a_1)

X_i	f_i	fX_i
1	5	5
2	2	4
3	1	3
4	2	8
5	0	0
6	3	18
	N=13	$38=fX_i$

a_2)

Clase	X_i	f_i	fX_i
1 a 2	1,5	7	10,5
3 a 4	3,5	3	10,5
5 a 6	5,5	3	16,5
		N=13	$37,5=fX_i$

Hemos partido de 13 puntuaciones correspondientes a una distribución de 6 categorías cuyo tamaño es la unidad. En la tercera columna (fX_i) aparecen los números de casos totales dentro de cada categoría. La suma de estos totales parciales, fX_i, es igual a 38. Pero si ahora agrupamos los mismos datos en tres categorías cuya anchura sea 2, en lugar de 1, tal como aparece en el apartado a_2) del cuadro, la columna de frecuencias totaliza el mismo número que en el caso anterior, pero no ocurre así con la columna del total de casos en cada categoría, cuya suma ya no es 38, sino 37,5. La diferencia entre ambos números se debe a que hemos calculado los totales parciales fX_i utilizando el valor medio de cada categoría. Precisamente la diferencia entre 38 y 37,5 es lo que se llama error de agrupamiento, y se produce porque los puntos medios de cada clase o categoría en el ejemplo no representan convenientemente el valor de los casos que se engloban en cada categoría.

Por tanto, al agrupar los datos, las categorías se calcularán con sumo cuidado, de forma que los valores medios de cada una de ellas refleje de la forma más exacta posible el valor de los casos en la categoría. Spiegel (1975, pág. 28) ofrece las dos siguientes reglas para formar las distribuciones de frecuencias y minimizar el error de agrupamiento: 1) determinar el mayor y el menor entre los datos registrados y así encontrar el *rango* (diferencia entre el mayor y el menor de los datos); 2) dividir el rango en un número conveniente de intervalos de clase de idéntico tamaño. Si ello no fuera posible, será preciso utilizar intervalos de clase de diferente tamaño e intervalos abiertos. El número de intervalos se pone generalmente entre 5 y 20, dependiendo de los datos de partida. Los intervalos se elegirán de forma que los puntos medios coincidan con datos realmente observados.

2.3.2.2. *Distribución porcentual*

Para calcular un porcentaje es preciso calcular previamente una proporción. La *proporción* de casos en una categoría dada es igual al número de casos en la categoría dividido por el número total de casos en la distribución. En una distribución de frecuencias de cinco categorías, en la que el número de casos en cada categoría fuese N_i y el número total de casos fuese N, la proporción de casos en cada categoría será N_i/N. Obviamente, el valor de una proporción no puede ser mayor que la unidad. Dado que

$$N_1+N_2+N_3+N_4+N_5=N$$

se tiene:

$$\frac{N_1}{N}+\frac{N_2}{N}+\frac{N_3}{N}+\frac{N_4}{N}+\frac{N_5}{N}=\frac{N}{N}=1$$

Por tanto, si se suman las proporciones de casos en todas las categorías, el resultado será la unidad. Se trata de una propiedad fundamental de las proporciones, y se puede generalizar a cualquier número de categorías.

Los *porcentajes* se obtienen a partir de las proporciones simplemente multiplicando por 100; de ahí que también se denominen *por ciento*. Al utilizar porcentajes, lo que se hace realmente es estandarizar según el tamaño, ya que se calcula el número de casos que habría en una categoría si el número total de casos fuera 100 y si la proporción en cada categoría no se alterase. Del mismo modo que la suma de las proporciones de una distribución dada es igual a la unidad, la suma de sus porcentajes será 100.

Si en lugar de los valores absolutos en una distribución de frecuencias se utilizan los correspondientes porcentajes, tendremos una *distribución porcentual*, que presenta algunas ventajas sobre la primera. Sobre todo, facilita la comparación, aparte de evitar una fuente importante de error. El porcentaje, que es en realidad una razón simple, se entiende fácilmente porque, tal como señalan Loether y McTavish (1974, pág. 54), tendemos en nuestra cultura a pensar en términos de partes de 100.

La distribución por edades anterior se puede escribir en términos de porcentajes del siguiente modo:

Edad (años)	f_i	%
De 18 a 20	10	8,33
De 21 a 25	14	11,66
De 26 a 35	23	19,16
De 36 a 45	20	16,66
De 46 a 60	29	24,16
De 61 a 75	24	20,00
	N=120	99,97

Para calcular el porcentaje de cada categoría se ha dividido cada f_i por N y se ha multiplicado por 100. Obsérvese que la suma de los porcentajes no es exactamente 100,0, debido a que sólo hemos tomado dos cifras decimales y no hemos redondeado el porcentaje resultante. Es aconsejable utilizar una sola cifra decimal, redondeándola de forma que si el número de la centésima es menor de 5 se mantiene el valor de la décima, pero si el número de la centésima es 5 o superior a 5 se incrementa en una unidad la cifra de las décimas. Realizando esta operación de redondeamiento, la anterior distribución porcentual quedaría del siguiente modo:

Edad (años)	%
De 18 a 20	8,3
De 21 a 25	11,7
De 26 a 35	19,2
De 36 a 45	16,7
De 46 a 60	24,2
De 61 a 75	20,0
TOTAL	100,1
	(120)

Ahora, la suma porcentual es 100,1, es decir, una décima superior a 100, por efecto de la operación de redondeamiento. Obsérvese también que el número que representa los casos totales N se ha puesto, entre paréntesis, debajo del 100. Esta práctica es habitual en la presentación de las tablas de distribuciones porcentuales, porque de este modo se indica la base real sobre la què se ha calculado el porcentaje.

Resulta conveniente señalar que, para calcular porcentajes, el valor de N ha de ser suficientemente elevado. Blalock (1960, pág. 28) señala el número 50 como el mínimo aproximado de casos que ha de contar una distribución para poder calcular los porcentajes. Si el número de casos es bastante inferior a 50, resulta más adecuado ofrecer el número real de casos en cada categoría en lugar de los porcentajes.

No siempre puede estar indicado la utilización de porcentajes para realizar comparaciones significativas y, en tal caso, convendrá operar con las cifras absolutas. Zeisel, que ha escrito quizá los capítulos más didácticos en el campo de la metodología de las ciencias sociales sobre el uso de los porcentajes, utiliza el siguiente ejemplo. Supongamos que queremos comparar dos empresas en términos de la variación anual de sus ventas. Supongamos que la empresa A ha aumentado en el último año su volumen de ventas de 1 a 2 millones de pesetas, lo que significa un aumento del 100 por 100; mientras que la empresa B ha pasado en sus ventas de 4 a 7 millones de pesetas, lo que significa un aumento del 75 por 100. Si comparamos las empresas A y B según sus cifras absolutas, B aventaja claramente a A, ya que sus ventas experimentaron un incremento de 3 millones de pesetas, mientras que la segunda experimentó una subida de sólo 1 millón. Sin embargo, si comparamos las dos empresas según sus incrementos porcentuales o relativos, la empresa A, con el 100 por 100, claramente supera a la empresa B, que sólo aumentó el 75 por 100. Para Zeisel, en caso de duda sobre la forma en que deben realizarse las comparaciones, «la consideración más general es presentar el aumento de forma que determine tan exactamente como sea posible el concepto que deseamos medir» (Zeisel, 1962, pág. 27).

Tampoco recomienda Zeisel el uso de porcentajes que excedan considerablemente de 100. Decir, por ejemplo, que los visitantes extranjeros en España aumentaron en la década de los sesenta un 1.200 por 100 sobre el número de visitantes en la década de los cincuenta puede producir una cifra impresionante, pero estadísticamente es un recurso muy pobre; resulta más correcto decir que el número de visitantes aumentó 12 veces en relación al período anterior.

2.3.2.3. *Distribución acumulada*

Una distribución acumulada se forma al indicar para cada categoría el número (o porcentaje) de casos que quedan por debajo del límite real superior de dicha categoría. Normalmente, se sigue la convención de crear distribuciones acumuladas, comenzando a acumular desde las categorías de orden inferior e ir así acumulando hasta N o 100 por 100, según se trate, respectivamente, de una distribución de frecuencias o una distribución porcentual. Para el caso de la distribución por edad que venimos utilizando, las dos distribuciones acumuladas quedarían del siguiente modo:

Edad (años)	*Frecuencia*	*Porcentaje*	*Frecuencia acumulada*	*Porcentaje acumulado*
De 18 a 20	10	8,3	10	8,3
De 21 a 25	14	11,7	24	20,0
De 26 a 35	23	19,2	47	39,2
De 36 a 45	20	16,7	67	55,9
De 46 a 60	29	24,2	96	80,1
De 61 a 75	24	20,0	120	100,1
	N = 120			

Así, para la categoría de 36 a 45 años, la frecuencia acumulada de 67, o el porcentaje de acumulado de 55,9 por 100, indican que el número o porcentaje de individuos con esa edad o menos es el que se indica.

Las distribuciones acumuladas son útiles en la comparación cuando se desea comparar la forma en que los casos se distribuyen a lo largo de una escala. Así, por ejemplo, al comparar los niveles de ingresos familiares en hogares españoles cuyo cabeza de familia pertenece a la clase social alta y media alta, o a la clase obrera, se obtienen los siguientes resultados:

TABLA 3

Distribuciones porcentuales acumuladas de los ingresos familiares, por clase social

Cantidad de pesetas mensuales	*Clase social alta y media-alta* (1)	*Clase social obrera* (2)	*Frecuencia acumulada* (1)	*Frecuencia acumulada* (2)
Más de 50.000	4	—	4	—
De 30.501 a 50.000	7	1	11	1
De 20.501 a 30.500	22	1	33	2
De 14.501 a 20.500	19	4	52	6
De 12.501 a 14.500	8	5	60	11
De 10.501 a 12.500	9	8	69	19
Menos de 10.000	31	81	100	100
	100	100		
	(279)	(1.126)		

FUENTE: FOESSA, 1970, pág. 563. Elaboración propia.

Las distribuciones acumuladas permiten una comparación más clara de las tremendas diferencias que, en materia de ingresos familiares, existían en los hogares españoles en el momento de realizar el estudio (finales de los años sesenta). Mientras que en los hogares cuyo cabeza de familia se identificaba con las clases sociales más altas el 69 por 100 disfrutaba de unos ingresos superiores a 10.500 pesetas, tal porcentaje era tan sólo del 19 por 100 en los hogares de familias obreras. De esta forma, vemos cómo los porcentajes acumulados permiten en una sola medida ofrecer los casos que se encuentran por debajo o por encima de unos niveles determinados.

2.3.3. *Percentiles*

El valor por debajo del cual queda un porcentaje determinado de casos es un *percentil*, y podemos representarlo por P_i, siendo i un valor que oscila entre 1 y 100. Así, el percentil 20 o P_{20} deja por debajo de su valor un 20 por 100 de casos.

El valor que divide a los datos en dos partes iguales, P_{50}, se llama también *mediana*. Por extensión, se puede hablar de aquellos valores que dividen a los datos en cuatro partes iguales. Estos valores, que podemos representar por Q_1, Q_2 y Q_3, se llaman primero, segundo y tercer *cuartil*, respectivamente; el valor de Q_2 es el valor que divide a los datos en dos partes iguales, y que se denomina mediana. De igual modo, los valores que dividen los datos en diez partes iguales se denominan *deciles*, y podemos representarlos por D_1, D_2, ..., D_9. Los resultados de muchas evaluaciones (tests) se presentan en forma de percentiles —el porcentaje de individuos que, en un determinado test, ha obtenido una puntuación igual o superior a un valor concreto.

Dos problemas de cálculo se presentan en relación a los percentiles. Cuando se desea calcular el rango de percentil de una puntuación determinada, hay que utilizar la siguiente fórmula:

$$\text{Rango de percentil de una puntuación dada} = \frac{\text{Lugar que ocupa la puntuación en la distribución}}{N} \times 100$$

En la distribución siguiente:

3
5
9
11
15
17
22

el rango de percentil del valor «11» será $= \frac{4}{7} \times 100 = 57{,}1$, ya que el valor 11 ocupa el cuarto lugar y $N=7$.

De manera inversa, se puede calcular el valor o puntuación correspondiente a un rango de percentil dado. Para ello se multiplicará el percentil por N y, a continuación, se buscará en la distribución el lugar que corresponde al número así calculado. Por ejemplo, en la distribución anterior, al percentil 70 le corresponde la puntuación 15, que ocupa el quinto lugar, ya que 5 es lo que resulta de redondear el número 4,9, que se obtiene al multiplicar 0,70 por N, que en este caso es 7.

El uso de percentiles resulta muy apropiado cuando se desea comparar, dadas una serie de distribuciones, unos grupos específicos, situados en un lugar dado de las distribuciones, con otros grupos situados en el mismo o diferente lugar. Murillo Ferrol hace un buen uso de esta lógica de la comparación cuando, al estudiar la distribución de las rentas en Andalucía, señala y denuncia que el incremento absoluto del volumen de las rentas no ha venido acompañado de un proceso de mejora en la redistribución de tales rentas entre todas las clases de población, ya que los pobres, mayoritarios, continúan percibiendo una proporción pequeña de las rentas, mientras que los ricos, que son muy pocos, reciben la parte más amplia de los ingresos. En la tabla 4 hemos reproducido las comparaciones porcentuales que realiza Murillo Ferrol, en base a la proporción de ingresos que corresponde al 20 por 100 más pobre de los hogares, al 20 por 100 más favorable y al 5 por 100 último de los más favorecidos. Esta utilización de los percentiles sirve mejor que otro algoritmo para evidenciar el desequilibrio existente en la distribución de las rentas. Así, y observando con más detenimiento los resultados de la tabla 4, se aprecia que, para el conjunto nacional, el 20 por 100 más pobre de la población recibe tan sólo el 6,8 por 100 de los ingresos totales, frente al 45,2 por 100 de ingresos que recibe el 20 por 100 más favorecido, o el 19,4 por 100 de ingresos que recibe el 5 por 100 más favorecido. Para las ocho provincias andaluzas, los resultados comparativos son parecidos a los de la media nacional, lo que revela un sistema de distribución económico muy injusto.

TABLA 4

Ingresos que corresponden a determinados grupos de la población

Provincias	*% de ingresos que correspon- de al 20 % más pobre de los hogares*	*% de ingresos que correspon- de al 20 % más favorecido*	*% de ingresos que correspon- de al 5 % último de los más favorecidos*
Almería	8,0	43,3	19,5
Cádiz	8,2	41,9	15,3
Córdoba	8,1	43,8	20,8
Granada	7,8	47,5	20,8
Huelva	8,5	42,0	17,1
Jaén	7,7	46,2	24,0
Málaga	8,5	38,0	15,1
Sevilla	7,5	46,4	20,7
España	6,8	45,2	19,4

FUENTE: MURILLO FERROL, F., «La distribución de la renta en Andalucía», *Anales de Sociología*, 4, 1968, pág. 40.

2.3.4. *Razón*

La *razón* * de un número A a otro número B se define como A dividido por B. La cantidad que precede a la palabra clave «a» se coloca en el numerador, mientras que el número que le sigue va al denominador. Así, si en un parlamento hay 160 diputados de izquierdas, 150 diputados de derechas y 80 diputados regionalistas, la razón de los diputados de izquierda a los diputados de derecha será 160/150, mientras que la razón de los diputados de izquierda y regionalistas a los diputados de derecha será (160+80)/150. Se puede, pues, utilizar un número compuesto tanto en el numerador como en el denominador, aunque el resultado se suele expresar de la forma numérica más simple posible, en este caso como 24/15.

Las proporciones y los porcentajes son un tipo especial de razón, en donde el denominador es el número total de casos y el numerador una fracción dada de dicho número, en el caso de la proporción, y esa misma fracción del número multiplicada por 100 en el caso del porcentaje. Pero, a diferencia de la proporción, la razón puede ser mayor que la unidad, como ocurre en el ejemplo de los diputados empleado anteriormente.

Las razones se suelen expresar también en términos de cualquier base que resulte conveniente para nuestros objetivos descriptivos. Una razón muy empleada en demografía es la de sexos, que se define como el número de varones de una población determinada dividido por el número de mujeres. Dado que el número total de varones es menor que el número de mujeres (aunque nacen más niños que niñas, la tasa de mortalidad masculina es mayor que la tasa de mortalidad femenina, por lo que entre la población adulta es mayor el número de mujeres que el de varones), la razón de los sexos será un número decimal, por lo que convencionalmente se suele multiplicar por 100, con lo que una razón de sexos de 94 indicará que hay 94 varones por cada 100 mujeres.

Cuando se utilizan bases mayores que 100, tales como 1.000, 10.000 o un millón, tenemos las *tasas*, que son otro tipo de razón. Las tasas se emplean cuando el uso de porcentajes arroja números decimales. Las tasas también se utilizan abundantemente en demografía y, en general, cuando se quiere disponer de indicadores sencillos referentes a la población general. Así, una tasa bruta de natalidad de 30 por 1.000 significa que se han producido 30 nacimientos por cada 1.000 habitantes.

Las *tasas de crecimiento relativo* son otro tipo muy utilizado de razón. Para calcular la tasa de crecimiento en un período de tiempo dado, se toma el incremento real durante el período y se divide por el tamaño que había al comienzo del período. Así, si la renta *per capita* en un país determinado ha pasado, en el período 1960-1970, de 1.500 a 2.000 dólares, la tasa de crecimiento relativo de la renta *per capita* será:

$$\frac{2.000-1.500}{1.500}=\frac{500}{1.500}=0{,}33$$

o, si se quiere expresar en términos porcentuales, del 33 por 100. Naturalmente, cuando la cantidad al final del período sea más reducida que

* Algunos tratadistas utilizan la palabra inglesa *ratio* para referirse al término razón, aunque resulta conveniente emplear este último.

al comienzo, la tasa resultante será negativa por reflejar un decrecimiento. También se pueden expresar las tasas de crecimiento relativo en relación a 1.000, 10.000 u otra cantidad que resulte conveniente con fines descriptivos y analíticos. En general, la tasa de crecimiento relativo se puede expresar como $(b - a/a) \cdot k$, siendo a y b las cantidades al principio y al final del período, respectivamente, y k la base que se decida utilizar, y que normalmente será una constante con ceros, del tipo de 1.000, 10.000, un millón, etc.

2.4. Técnicas básicas de representación gráfica

Los resultados de las investigaciones estadísticas se suelen representar muchas veces gráficamente, con el fin de obtener un panorama más intuitivo y directo de los mismos. Aunque son muchos los recursos gráficos que los sociólogos utilizan para ofrecer una visión directa y simple de sus investigaciones cuantitativas, aquí vamos a referirnos, para comenzar, a las representaciones gráficas utilizadas en estadística para el caso de las distribuciones de frecuencia. Los *histogramas*, los *polígonos* y las *ojivas* son tales representaciones, y junto con la *línea de grafos*, constituyen los procedimientos gráficos básicos más utilizados en el campo de la estadística.

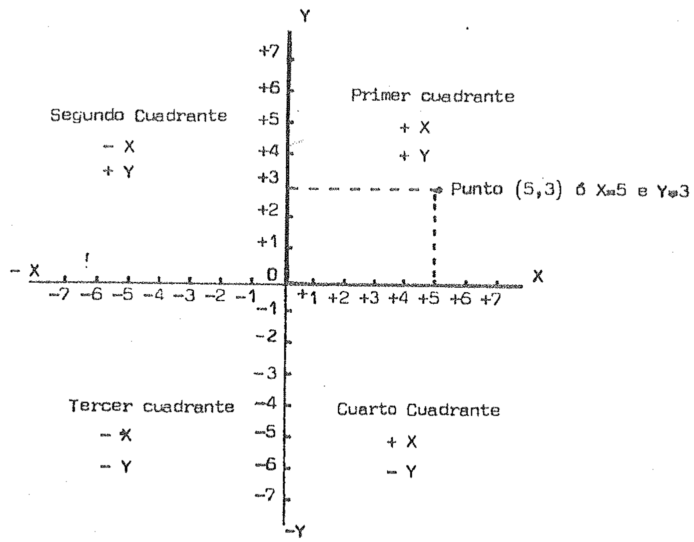
En toda representación gráfica se encuentra subyacente la idea de un *sistema de referencias* o *sistema de coordenadas*. El sistema de coordenadas más usual en las representaciones gráficas consiste en dos líneas, o «dimensiones», perpendiculares que forman el sistema de Coordenadas Cartesianas —en honor del filósofo René Déscartes (1596-1650), que fue el primero en combinar el álgebra con el análisis gráfico—. Como es sabido, la línea o eje vertical se llama *ordenada* o eje de las Y, y la línea o eje horizontal se denomina *abscisa* o eje de las X. Ambos ejes dividen al plano en cuatro *cuadrantes*, y el punto donde se cruzan ambos ejes se denomina *origen* o *punto cero*, ya que las escalas numéricas parten del origen en las cuatro direcciones. Las puntuaciones que parten del origen hacia arriba por el eje Y y, a la derecha, por el eje X son positivas, mientras que las puntuaciones que parten del origen hacia abajo por el eje X y, a la izquierda, por el eje Y son negativas. Dado que la mayoría de las mediciones en sociología se realizan en escalas que parten desde cero sólo en la dirección positiva, el cuadrante primero es el que se suele necesitar preferentemente, por lo que en las representaciones gráficas se omiten con frecuencia el resto de los cuadrantes y sólo se representa el primer cuadrante (ver fig. 1).

2.4.1. *Histogramas*

Un histograma, o histograma de frecuencias, consiste en la representación de una distribución de frecuencias o porcentual, en la que la fre-

FIGURA 1

Sistema referencial de Coordenadas Cartesianas



cuencia (o porcentaje) de casos en cada categoría se representa mediante un rectángulo que tiene su base sobre un eje horizontal (el eje X), con centro en el punto medio de la categoría y cuya anchura es igual al tamaño del intervalo de dicha categoría. La altura del rectángulo será proporcional a la frecuencia (o porcentaje) de casos que se incluyen en la categoría. Es decir, el área del rectángulo será proporcional a la correspondiente frecuencia (absoluta o relativa).

Supongamos que hemos elaborado la distribución de frecuencias que se incluye en la tabla 6 a partir de los datos que recoge el Censo de Población de España de 1970, referente al tamaño de las familias españolas.

TABLA 6

Distribución de las familias según el número de miembros

Número de miembros. Categoría	*Punto medio*	*Frecuencia (en miles)*
0,0 - 3	1,5	3.959
3,1 - 6	4,5	4.094
6,1 - 9	7,5	706
9,1 - 12	10,5	94
	Total	8.853

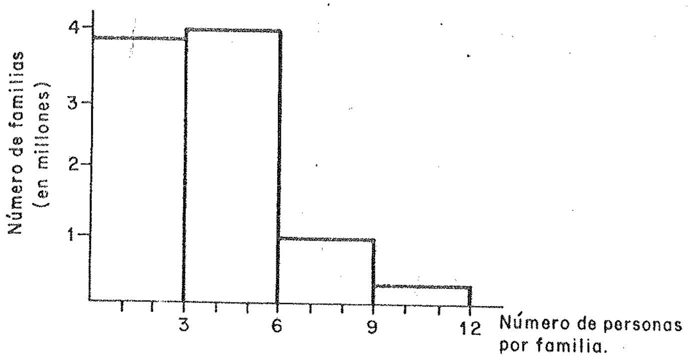
Fuente: *Censo de Población 1970.*

Convencionalmente, hemos cerrado el límite superior de la categoría o clase más alta por medio del número 12, con el fin de tener intervalos iguales. De este modo, las familias españolas se han clasificado en cuatro categorías, según las formen tres o menos miembros, cuatro a seis, siete a nueve o diez a doce miembros. Realmente, no existen familias que tengan 3,1, 6,1 o 9,1 personas. Se trata simplemente de una convención aritmética que hemos adoptado para fijar unos límites de intervalo que sean mutuamente excluyentes.

Pues bien, el histograma de frecuencias correspondiente a la anterior distribución será el que se incluye en la figura 2, y cuyos rectángulos tendrán como base intervalos de tres puntos y como alturas las correspondientes frecuencias.

FIGURA 2

Histograma del tamaño de las familias españolas (1970)



Si en lugar de disponer de las cifras absolutas tuviéramos los porcentajes de familias correspondientes a cada una de las categorías, el histograma resultante, que ahora sería un histograma de frecuencias porcentual, se construiría de idéntico modo.

Los histogramas se utilizan con ciertas variaciones, según que los datos que forman la distribución de frecuencias se encuentren medidos a nivel nominal, ordinal o de intervalo. Así, si la variable es nominal, los rectángulos del histograma se separan ligeramente unos de otros, con el fin de visualizar que se trata de categorías diferentes.

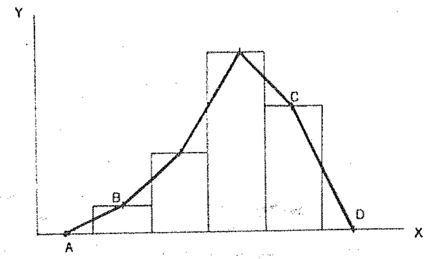
Cuando se trata de variables ordinales, en las que no se definen distancias iguales, también se suelen separar ligeramente los rectángulos para destacar tal hecho, aunque hay quien prefiere mantener las columnas juntas con el fin de conservar la impresión de «escalera» del histograma. También se suele adoptar una base estándar de amplitud constante, a pesar de que las distancias no están definidas. Vemos, pues, que

las variables ordinales se pueden tratar como si fueran de intervalo, a efectos de construir el histograma, siempre y cuando se tengan presente las convenciones adoptadas para poder realizar la representación. Con todo, el histograma está más indicado cuando las variables se encuentran medidas a nivel de intervalo.

2.4.2. *Polígonos*

El polígono de frecuencias (o de porcentajes) es una figura que se cierra al unir los puntos medios de cada intervalo, a una altura proporcional a la frecuencia (o porcentaje) de dicho intervalo. La unión de tales puntos constituye un segmento rectilíneo que, al prolongarlo por los extremos hasta cortar al eje X, constituye un polígono de frecuencias. El área que queda por debajo del polígono de frecuencias es igual al área contenida dentro del correspondiente histograma.

En el siguiente gráfico se observa la construcción de un polígono a partir del correspondiente histograma:

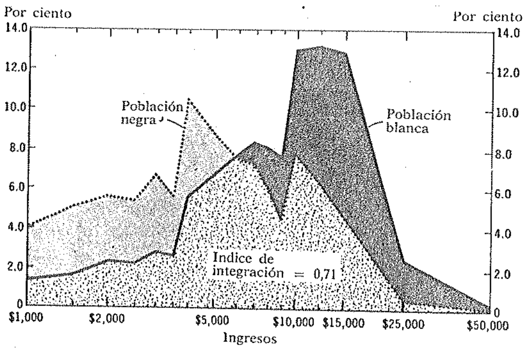


Obsérvese cómo se ha prolongado el segmento rectilíneo BC, constituido al unir los puntos superiores de cada intervalo, mediante el trazado de los segmentos BA y CD hasta los puntos medios de categoría inferior y superior inmediatos, y que corresponden a la clase de frecuencias cero. De este modo, el polígono queda cerrado y el área que contiene es igual al área de la suma de los rectángulos.

A veces resulta de interés representar los polígonos de varios grupos considerados conjuntamente. De este modo se pueden observar las áreas en donde las distribuciones correspondientes coinciden o se separan. Weitzman (1970, pág. 9), al estudiar las distribuciones de los ingresos de las familias de población blanca y negra en los Estados Unidos, utilizó el área de coincidencia de ambas distribuciones como una medida de integración, que la obtuvo mediante el cálculo del porcentaje o proporción del área de coincidencia de ambas distribuciones (ver fig. 3).

FIGURA 3

Polígonos de frecuencias porcentuales correspondientes a las distribuciones de ingresos en familias de población blanca y negra en los Estados Unidos.



FUENTE: M. S. WEITZMAN, 1970, pág. 9.

En este ejemplo, que hemos tomado del análisis gráfico realizado por Weitzman, se produce una coincidencia del 71 por 100 del área de ambos polígonos de frecuencias. Es decir, que el índice de integración de los ingresos familiares entre la población blanca y negra es igual a 0,71. Una segregación completa, del 0,00, vendría dada por una representación gráfica en la que no se produjeran coincidencias, mientras que una integración completa, del 1,00, se produciría si coincidieran ambas curvas.

2.4.3. *Ojivas*

Las ojivas son polígonos de frecuencias acumuladas. El primer punto de dicho polígono vendrá dado por el límite real inferior del primer intervalo. A continuación, en la vertical sobre el límite real superior de cada intervalo, y a una altura proporcional a la frecuencia (o porcentaje) acumulada de dicho intervalo, dibujamos un punto. Uniendo mediante un segmento rectilíneo cada dos puntos consecutivos se obtiene, para el conjunto de todos los pares de puntos unidos, el polígono de frecuencias acumuladas u ojiva.

Supongamos que tenemos la distribución de frecuencias de los ingresos mensuales de un grupo de 100 trabajadores y obtenemos la co-

rrespondiente distribución de frecuencias acumulada, tal como aparece en la tabla 6.

TABLA 6

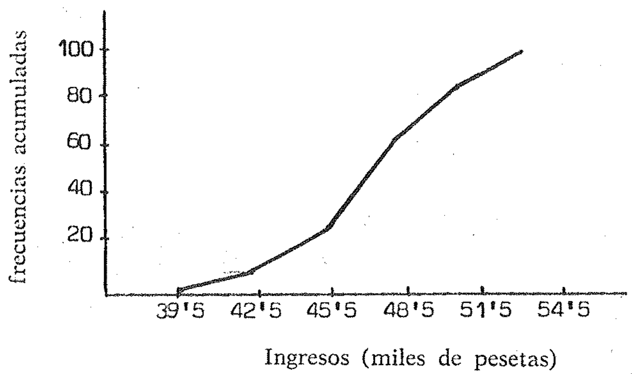
Distribución de frecuencias absoluta y acumulada de los ingresos mensuales de un grupo de 100 trabajadores

Distribución de frecuencias		*Distribución de frecuencia acumulada*	
Ingresos (miles de pesetas)	*N.º de trabajadores*	*Ingresos (miles de pesetas)*	*N.º de trabajadores*
40 a 42	5	Menos de 39,5	0
43 a 45	18	Menos de 42,5	5
46 a 48	40	Menos de 45,5	23
49 a 51	29	Menos de 48,5	63
52 a 54	8	Menos de 51,5	92
		Menos de 54,5	100
TOTAL	100		

Para construir el polígono de frecuencias acumuladas u ojiva correspondiente a la distribución anterior representamos los límites reales en el eje *X* y las frecuencias acumuladas en el eje *Y*, tal como aparece en la figura 4.

FIGURA 4

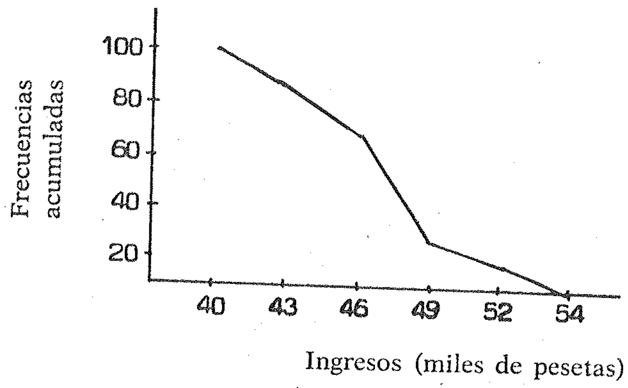
Ojiva correspondiente a la distribución de frecuencias acumuladas de los ingresos mensuales de 100 trabajadores



Algunas veces resulta más conveniente considerar una distribución de frecuencias acumuladas de todos los valores mayores o iguales al límite real inferior de cada intervalo de clase. Por lo que se refiere al ejemplo anterior, podríamos considerar los ingresos superiores a 39.500 pesetas o más, 42.500 pesetas o más, etc. Así se obtendría una *distribución acumulada «o más»*, a diferencia de la anterior, que se denomina *distribución acumulada «menos de»*. El paso de un tipo de distribución a otra es bien sencillo, a partir de la distribución de frecuencias absolutas, como se observa a continuación:

Distribución de frecuencias		*Distribución de frecuencias acumulada «menos de»*		*Distribución de frecuencias acumulada «o más»*	
Ingresos (miles de ptas.)	*Núm. de trabajadores*	*Ingresos (miles de ptas.)*	*Núm. de trabajadores*	*Ingresos (miles de ptas.)*	*Núm. de trabajadores*
40 a 42	5	Menos de 39,5 ...	0	40 o más	100
43 a 45	18	Menos de 42,5 ...	5	43 o más	95
46 a 48	40	Menos de 45,5 ...	23	46 o más	77
49 a 51	29	Menos de 48,5 ...	63	49 o más	37
52 a 54	8	Menos de 51,5 ...	92	52 o más	8
		Menos de 54,5 ...	100	54 o más	0
TOTAL	100				

La representación gráfica de una distribución de frecuencias acumuladas del tipo «o más» dará lugar a una ojiva de pendiente inversa a la de la ojiva resultante de representar gráficamente una distribución de frecuencias acumuladas del tipo «menos de», como se observa en el siguiente gráfico:



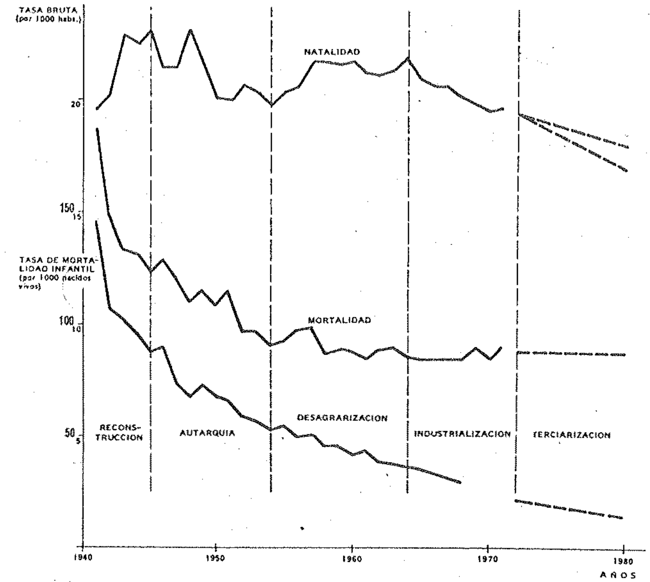
2.4.4. *Línea de grafos*

Otra de las técnicas de representación gráfica es la línea de grafos, que muestra el valor de una variable dependiente (que se representa a lo largo del eje Y) para cada valor de las diferentes categorías de otra variable, normalmente utilizada como variable independiente (que se representa a lo largo del eje X). Los puntos representados en el primer cuadrante se unen mediante una línea continua hasta el último de los puntos representados. La línea no se cierra sobre el eje X, como se hace en la representación de los polígonos, ya que el área que queda por debajo de la línea de grafos no tiene ningún significado espacial, como ocurre con los histogramas y polígonos. Más bien, lo que le interesa al investigador de la línea de grafos es la forma que adopta, la pendiente que toma al crecer o decrecer la línea y, en el caso de representar sobre el mismo cuadrante dos o más líneas, las semejanzas y diferencias que presentan.

Como ejemplo de esto último obsérvese la evolución de las tasas de natalidad, mortalidad y mortalidad infantil en España para el período 1941-1971, tal como han sido estudiadas por Amando de Miguel (1974, pá-

FIGURA 5

Líneas de grafos correspondientes a las tasas de natalidad, mortalidad y mortalidad infantil en España (1941-1971)



FUENTE: Amando DE MIGUEL, *Manual de estructura social de España*, Madrid, Tecnos, 1974, pág. 45.

gina 45). En el eje Y se han representado los valores de las tasas de natalidad y mortalidad, y en el eje X la variable años, haciendo coincidir el año 1940 con el origen o punto cero. Al representar las tres líneas de grafos en el mismo gráfico, el autor del análisis gráfico realiza una serie de consideraciones que le vienen dadas por la naturaleza comparativa de los datos. Además, al cubrir los datos un período de treinta años, Amando de Miguel relaciona los cambios que se observan en la evolución de las tasas con las fases experimentadas por la sociedad española, desde el punto de vista de su desarrollo económico. Además, el autor realiza una proyección al año 1980 de la evolución de las correspondientes tasas, continuando las tendencias que manifiestan las líneas de grafos.

2.5. Otras técnicas de representación gráfica

Aparte de las técnicas anteriores, los sociólogos utilizan otras técnicas que permiten realizar análisis muy útiles de los datos o una representación más asequible de los resultados. La pirámide de población, el gráfico rectangular, el gráfico de sectores y el gráfico triangular son cuatro de las técnicas gráficas más ampliamente usadas en sociología.

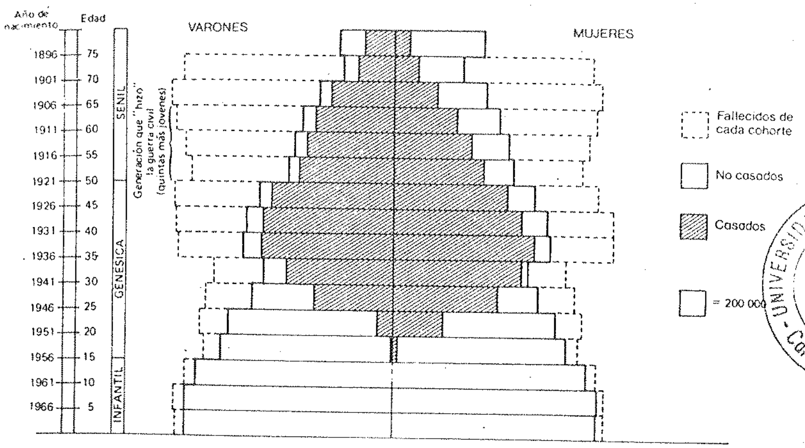
La *pirámide de población*, ampliamente utilizada en demografía, ofrece un diseño un poco más complejo de lo que llevamos visto hasta ahora, aunque en cierto modo refleja básicamente las ideas gráficas del histograma. En la figura 6 se ofrece un ejemplo de pirámide de la población en España según los datos del Censo de 1970, confeccionada por Amando de Miguel. Cada rectángulo en el gráfico representa el porcentaje de población en una categoría específica de edad y sexo. Los datos para varones se representan en la parte izquierda de la pirámide, mientras que los datos referentes a la población femenina aparecen en la parte derecha. Los rectángulos representan la proporción de población en grupos de cinco años, desde las edades más jóvenes, que forman la base de la pirámide, hasta las edades más avanzadas, que forman la cúspide.

En la pirámide de población anterior, el autor no sólo ha representado los porcentajes de población en cada categoría de sexo y edad, sino que también, para la población adulta, ha calculado la proporción de población casada de la que no está. Además, y en líneas de trazos, se ha representado la pirámide de población hipotética caso de que toda la población nacida en cada período no hubiera fallecido; así, por diferencia entre la pirámide real y la pirámide hipotética, se obtiene una visión directa de la proporción de personas fallecidas en cada categoría de edad.

Esta es precisamente la principal cualidad de la pirámide de población, al ofrecer directa e intuitivamente la distribución global de la po-

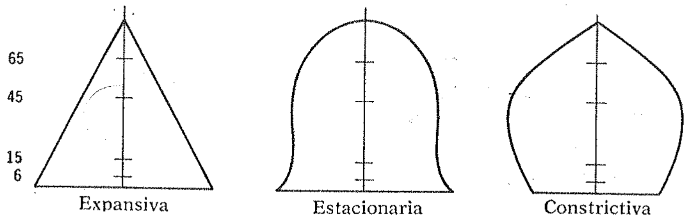
FIGURA 6

Pirámide de población en 1970



FUENTE: Amando DE MIGUEL, *La pirámide social española*, Madrid, Ariel, 1977, página 167.

blación y permitir comparaciones entre diferentes tipos de población. Al comparar las pirámides de población entre países se suelen distinguir tres formas principales de pirámides: las que ofrecen las poblaciones expansivas, las poblaciones estacionarias y las poblaciones constrictivas. Sus formas, idealizadas, son las siguientes:



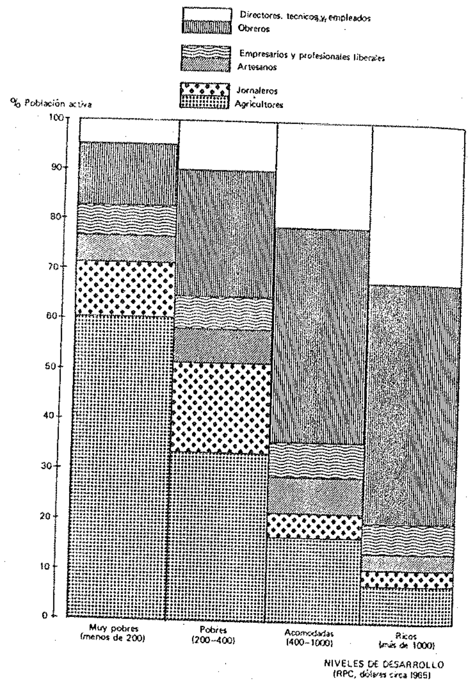
Según se asemejen las pirámides de población obtenidas para cada país a uno u otro modelo, así se podrá caracterizar el tipo de población. Por lo que se refiere a la pirámide de población española de 1970, su perfil se encuentra a medio camino entre el modelo expansivo y el estacionario. Esto indica que la población española, en la medida que va experimentando un descenso en las tasas de natalidad, va dejando atrás el modelo expansivo y se va acercando al modelo estacionario.

El *gráfico rectangular* es una variación del histograma, utilizado bien para representar variables nominales o bien para destacar categorías específicas de variables. En la figura 7 se ha representado la estructura ocupacional de diversos países, agrupados según los niveles de desarrollo alcanzados. En el gráfico se distinguen seis tipos de ocupaciones, y la altura de cada rectángulo representa el porcentaje de población activa en cada ocupación. De este modo, se puede observar a simple vista la rápida disminución de la población agraria según se pasa de un nivel de desarrollo inferior a uno superior, compensado por el crecimiento de la población obrera industrial y de los directivos, técnicos y empleados.

Un *gráfico de sectores* aparece reproducido en la figura 8, en donde un círculo representa el total de una característica, en este caso una muestra de científicos españoles. Al clasificarlos según su nivel de religiosidad, los sectores circulares correspondientes representan con sus

FIGURA 7

Estructura ocupacional de los países agrupados según el nivel de desarrollo

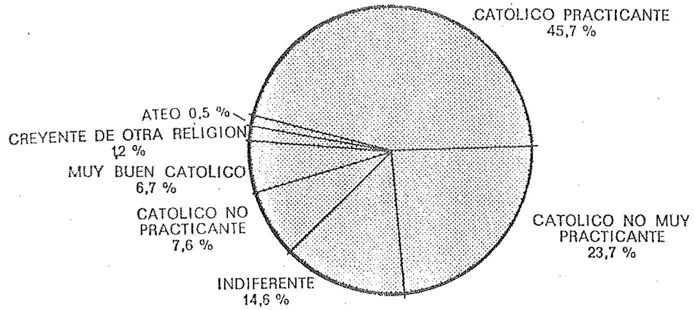


FUENTE: Amando DE MIGUEL, *Manual de estructura social de España*, Madrid, Tecnos, 1974, pág. 362.

áreas los porcentajes totales de científicos en cada categoría. En una representación sectorial, pues, el círculo se divide en sectores cuyo ángulo refleja el porcentaje del total para cada categoría. Como un círculo tiene 360°, al dividirlo en 100 partes, cada 3,6° representa una unidad porcentual del total. Así, los científicos que se declaran católicos practicantes ocuparán un sector cuyo ángulo será de 164,5° (45,7% × 3,6° = = 164,5%).

FIGURA 8

Nivel subjetivo de religiosidad de una muestra de científicos españoles



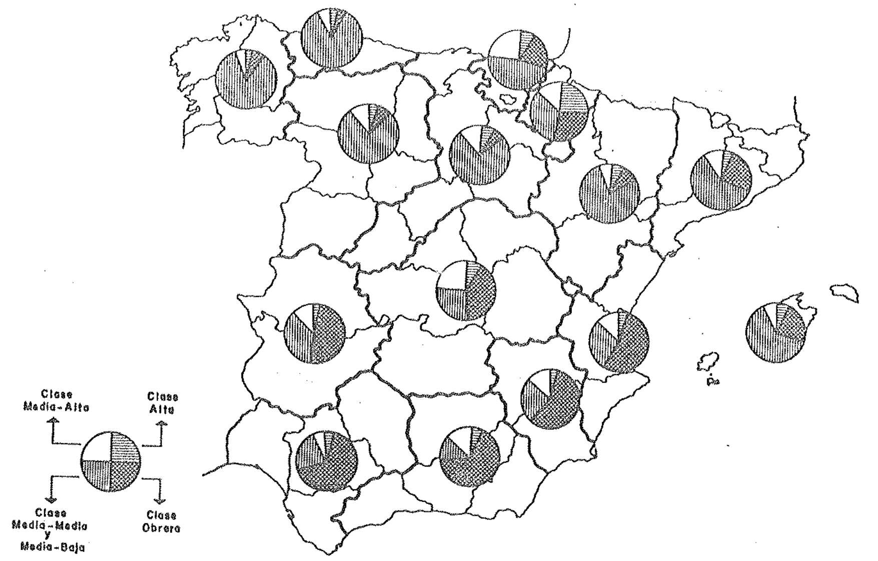
FUENTE: Pedro GONZÁLEZ BLASCO, *El investigador científico en España*, Madrid, C.I.S., 1980, pág. 161.

Con frecuencia, cada sector se raya de forma diferente con objeto de permitir un mayor contraste al comparar cada categoría. También se suelen utilizar los gráficos de sectores para comparar categorías de población para diversas áreas geográficas. De este modo, representando en un mapa dividido en regiones o áreas geográficas los diferentes círculos subdivididos según los valores que en cada una de ellas toman las diversas categorías de población, se consigue en una sola representación gráfica introducir un gran volumen de información, que permite obtener una buena imagen del conjunto. Esto es lo que se ha hecho en el mapa que se reproduce en la figura 9, en el que se ha representado sobre cada región española un gráfico circular, subdividido cada uno de ellos en cuatro segmentos que representan los correspondientes porcentajes de población activa agraria, diferenciada en cuatro clases o estratos sociales.

El *gráfico triangular* es una especie de diagrama de dispersión en el que cada caso o individuo se localiza mediante un punto o señal en el espacio del gráfico, de tal modo que se pueden examinar los conglomerados y las distancias entre los puntos referentes a variables determinadas. En el caso del gráfico triangular, los puntos se localizan en un gráfico que tiene la forma de un triángulo equilátero.

FIGURA 9

Estratificación social agraria por regiones



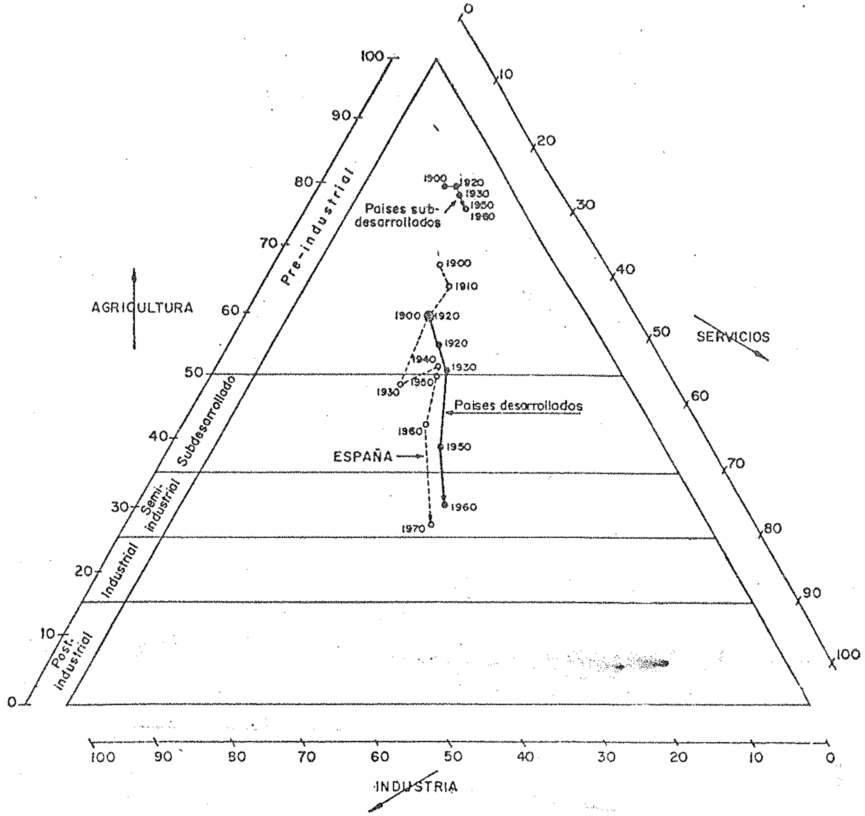
FUENTES: Manuel GARCÍA FERRANDO, «Estratificación social en el campo español», *Revista de Estudios Agrosociales*, 102, 1978, pág. 21.

Se suele utilizar en aquellas situaciones en que una misma variable consta de tres categorías significativas (o, al menos, se pueden agrupar en tres categorías), y en donde un caso particular puede ser caracterizado en términos de un porcentaje en cada categoría, de tal forma que el total para cada caso será 100 por 100. Habitualmente, los casos que se representan en el gráfico triangular son grupos, tales como «población agraria», «población industrial» y «población de servicios», o una población que puede ser caracterizada en función de un rasgo que adopta tres categorías, y que totaliza para cada caso 100 por 100, como «favorable», «desfavorable» y «no opina», en relación a un tema determinado en una encuesta de opinión.

El papel en el que se representa el gráfico triangular, y que se suele vender comercialmente, contiene un triángulo equilátero en el que en cada lado se representa una escala que va de un vértice a otro, en un recorrido de 100 unidades porcentuales. Un punto que se encuentre a un tercio del camino de cada vértice cae exactamente en el medio del gráfico, en un punto que vale el 33,3 por 100 en cada una de las tres

FIGURA 10

Distribución sectorial de la población activa en los países desarrollados, subdesarrollados y España (1900-1970)



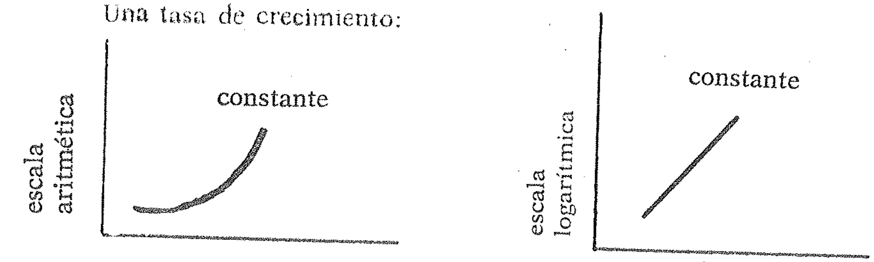
FUENTE: FOESSA, 1970, pag. 123.

escalas. En la figura 10 aparece representada triangularmente la evolución de la distribución sectorial de la población activa en los países desarrollados, subdesarrollados y en España para el período 1900-1970. De este modo se ha conseguido introducir en un solo gráfico una gran amplitud de información para un período de tiempo dilatado, permitiendo una comparación visual muy sencilla e intuitiva de las correspondientes evoluciones de la población activa para las tres categorías de países.

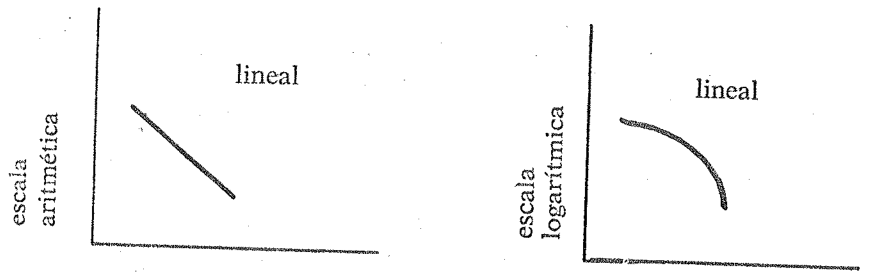
La última técnica gráfica que vamos a presentar en este capítulo es el *gráfico semilogarítmico*, que consiste en un gráfico representado en unas coordenadas rectangulares, similares a las coordenadas cartesianas, pero en las que en el eje *Y* se representa, en lugar de valores aritméticos, como en el eje X, los logaritmos de los números. Se distingue la representación logarítmica porque los intervalos que se señalan en el eje Y son de diferente amplitud, mientras que los intervalos en el eje X son constantes. El papel semilogarítmico también se vende comercialmente, por lo que el investigador no tiene que calcular los logaritmos de cada número que representa en el eje Y, ya que el papel trae marcados los valores correspondientes a los logaritmos de los números.

Este tipo de representación es particularmente útil en los análisis de líneas de tendencia, porque las diferencias numéricas iguales entre los logaritmos indican tasas de cambio iguales. Las siguientes figuras ilustran las diferencias existentes en una representación gráfica mediante una escala aritmética y una escala logarítmica:

Una tasa de crecimiento:



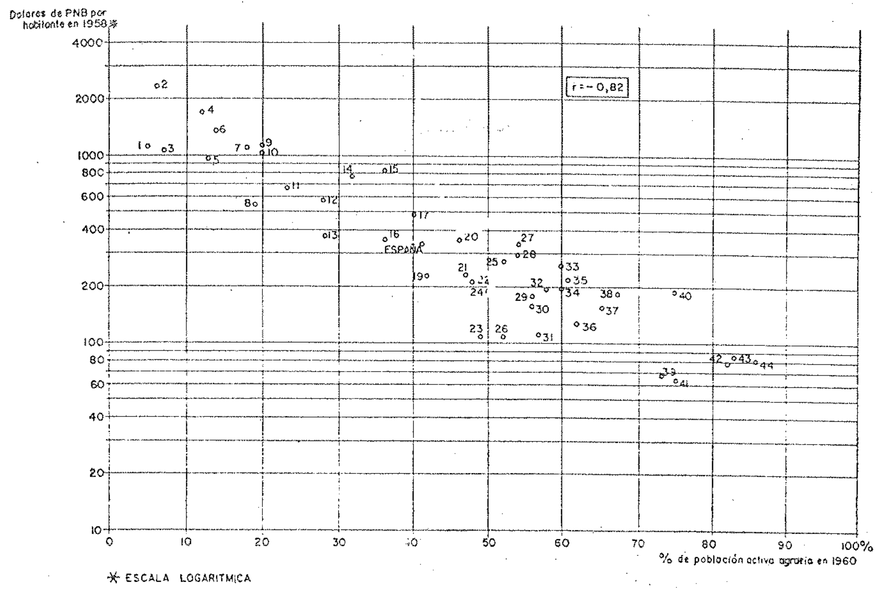
Una tasa de crecimiento:



Aparte del estudio de líneas de tendencias, la representación semilogarítmica se utiliza siempre que dispongamos de unos datos cuyos intervalos tienen un recorrido tan amplio que no cabrían en el papel en una representación gráfica de tipo aritmético. Al ser el valor del logaritmo de un número mucho menor que éste, el papel semilogarítmico permite representar en el mismo gráfico valores muy dispares. En la figura 11 se reproduce la representación gráfica de la correlación entre el producto nacional bruto por habitante y la proporción de población activa agraria, para varios países, utilizando una escala semilogarítmica, pues de este modo se ha podido representar conjuntamente países cuyos valores del PNB por habitante oscilan entre 70 y 2.500 dólares.

FIGURA 11

Relación entre el PNB por habitante y la proporción de población activa agraria, para varios países (circo 1960)



FUENTE: FOESSA, 1970, pág. 101.

Además de las aquí expuestas, existen otras técnicas gráficas que son también utilizadas por los científicos sociales, a veces no tanto con fines analíticos, sino más bien para ofrecer al público una imagen asequible de los resultados de sus investigaciones. No obstante, tales representaciones gráficas suelen ser variaciones de alguna de las técnicas básicas que hemos visto anteriormente, y con cuyo conocimiento resultan fácilmente interpretables el resto de ellas.

En el siguiente capítulo volvemos a ocuparnos de las distribuciones univariables, pero lo vamos a hacer por medio de la utilización de unos pocos números índices que resumen estadísticamente las características globales de las distribuciones.

2.6. Terminología

Se recomienda la memorización y comprensión del significado de cada uno de los términos y conceptos siguientes:

— Tipos de comparación.
— Operaciones básicas de comparación.
— Organización de los datos.
— Distribución de frecuencias.
— Distribución porcentual.
— Distribución acumulada.
— Clases o categorías.
— Intervalo de clase o categoría.
— Frecuencia de clase.
— Límites de clase: límite superior y límite inferior.
— Límites reales o verdaderos de clase o categoría.
— Amplitud, tamaño o longitud de clase o categoría.
— Punto medio de la clase o categoría.
— Error de agrupamiento.
— Rango.
— Proporción.
— Porcentaje.
— Percentil, cuartil, decil.
— Razón.
— Tasa, tasa de crecimiento relativo.
— Sistema de referencias o Sistema de coordenadas.
— Histograma.
— Polígono de frecuencias.
— Ojiva.
— Línea de grafos.
— Gráfico rectangular.
— Gráfico de sectores.
— Gráfico triangular.
— Gráfico semilogarítmico

EJERCICIOS

1. La variación de la población española provincial que se ha producido en el período 1940-1975, ha sido la que sigue, tomando la población de 1940 como base 100:

Alava, 211; Albacete, 88; Alicante, 175; Almería, 108; Avila, 80; Badajoz, 86; Baleares, 155; Barcelona, 227; Burgos, 92; Cáceres, 83; Cádiz, 159; Castellón, 132; Ciudad Real, 91; Córdoba, 94; Coruña (La), 118; Cuenca, 67; Gerona, 137; Granada, 100; Guadalajara, 68; Guipúzcoa, 206; Huelva, 109; Huesca, 93; Jaén, 86; León, 107; Lérida, 117; Logroño, 109; Lugo, 79; Madrid, 272; Málaga, 136; Murcia, 123; Navarra, 131; Orense, 90; Oviedo, 131; Palencia, 86; Palmas (Las), 221; Pontevedra, 129; Salamanca, 90; S. C. Tenerife, 191; Santander, 125;

Segovia, 80; Sevilla, 143; Soria, 65; Tarragona, 143; Teruel, 67; Toledo, 97; Valencia, 154; Valladolid, 136; Vizcaya, 225; Zamora, 77; Zaragoza, 135. Total España, 139.

Agrupar las provincias en categorías que sean sociológicamente significativas en relación a la tasa de variación de la población.

2. La población de los países europeos era, en 1983, la siguiente (en millones de personas):

Europa del Norte ...	82,0	*Europa Occidental*	155,0
Dinamarca	5,1	Alemania Federal	61,5
Finlandia	4,8	Austria	7,6
Irlanda	3,5	Bélgica	9,9
Islandia	0,2	Francia	54,6
Noruega	4,1	Luxemburgo	0,4
Reino Unido	56,0	Países Bajos	14,4
Suecia	8,3	Suiza	6,5
Europa Oriental	111,0	*Europa del Sur*	141,0
Alemania Oriental ...	16,7	Albania	2,9
Bulgaria	8,9	España	38,4
Hungría	10,7	Grecia	9,9
Polonia	36,6	Italia	56,3
Rumania	22,7	Malta	0,4
Checoslovaquia	15,4	Portugal	9,9
		Yugoslavia	22,3

Calcular los porcentajes que representan las poblaciones de cada país en relación al total europeo y al total del área geográfica a la que pertenecen.

3. En una encuesta sobre victimización, las 560 personas que respondieron afirmativamente a la pregunta: ¿ha sido víctima de algún delito o de algún intento de delito a lo largo de su vida?, se distribuyeron del siguiente modo de acuerdo con la edad:

Edad (años)	f_i
16-25	100
26-35	130
36-45	120
46-55	120
56-65	90

Desarrollar la distribución, calculando lo que sigue:

a) Frecuencias relativas y frecuencias acumuladas «menos de» y «o más».
b) Porcentajes y porcentajes acumulados.

4. Al visitar 84 hogares de un barrio, un encuestador encontró los siguientes números de miembros que viven en cada hogar:

3	5	6	1	2	4	4	3	5	1	3	2
4	5	3	6	4	5	3	4	8	7	4	3
4	5	4	2	6	7	1	3	2	4	3	4
6	7	4	5	3	4	7	4	3	5	6	3
7	2	9	10	3	2	1	3	2	3	4	4
6	7	1	2	3	4	7	3	4	6	1	3
8	4	3	2	3	9	3	6	2	2	4	5

a) Construir una distribución de frecuencias y una distribución acumulada, agrupando los datos en intervalos que sean sociológicamente significativos.
b) A partir de tales distribuciones, dibujar un histograma, un polígono de frecuencias y una ojiva.

5. Supóngase que los siguientes números representan los ingresos mensuales (en miles de pesetas) de una muestra de residentes en una comunidad determinada:

68	54	78	150	75	84	175	70	71	53	91	66
76	45	61	87	103	95	108	100	85	89	87	72
65	96	88	200	100	120	105	66	97	136	119	93
82	100	140	78	99	138	87	100	88	143	106	106
112	120	92	205	95	68	90	93	118	75	87	140
90	86	110	66	80	135	75	115	90	78	93	185

a) Construir una distribución de frecuencias y una distribución acumulada, agrupando los datos en intervalos que sean sociológicamente significativos.
b) A partir de tales distribuciones, dibujar un histograma, un polígono de frecuencias y una ojiva.

6. En 1960 la población urbana en España alcanzaba la cifra de 17.363.790 habitantes, y en 1970 llegó a la cantidad de 22.576.000. También en 1960 la población rural era 4.440.868, mientras que en 1970 descendió a 3.737.000. Calcular las tasas de crecimiento relativo (expresadas porcentualmente) para la población rural y para la población urbana, en el período considerado.

7. En una comunidad, el número de varones es 45.712 y el de mujeres es 47.523, ¿cuál es la razón de los sexos en dicha comunidad?

BIBLIOGRAFIA

ANDRÉS ORIZO, Francisco: *Cambio socio-cultural y comportamiento económico*, Madrid, Centro de Investigaciones Sociológicas, 1979.

BLALOCK, Hubert M.: *Social Statistics*, New York, McGraw-Hill, 1960.

DE MIGUEL, Amando: *Manual de estructura social de España*, Madrid, Tecnos, 1974.

DE MIGUEL, Amando: *La pirámide social española*, Barcelona, Ariel, 1977.

DURKHEIM, Emile: *Las reglas del método sociológico*, Buenos Aires, La Pléyade, 1972.

FOESSA: *Estudios sociológicos sobre la situación social de España 1975*, Madrid Euramérica, 1976.

GARCÍA FERRANDO, Manuel: «Estratificación social en el campo español», *Revista de Estudios Agrosociales*, 102, 1978, págs. 7-31.

GONZÁLEZ BLASCO, Pedro: *El investigador científico en España*, Madrid, C.I.S., 1980.

LOETHER, H. J., y D. G. MCTAVISH: *Descriptive Statistics for Sociologists*, Boston, Allyn and Bacon, 1974.

MURILLO FERROL, Francisco: «La distribución de la renta en Andalucía», *Anales de Sociología*, 4, 1968.

SPIEGEL, Murray E.: *Estadística*, Madrid, Ediciones de la Colina, S. A., 1975.

WEITZMAN, Murray S.: «Measures of Overlap of Income Distributions of White and Negro Families in the United States», *Technical Paper*, 22, Washington D.C., U.S. Bureau of the Census, 1970.

STOUFFER, Samuel A.: «Some observations on Study Design», *American Journal of Sociology*, 40, 950a, págs. 355-361.

ZEISEL, Hans: *Dígalo con números*, México, F.C.E., 1962.