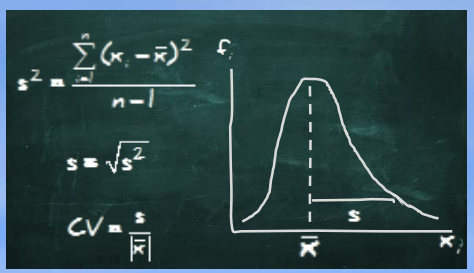
METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli



UAB
Universitat Autònoma de Barcelona

METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Pedro López-Roldán
Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (http://quit.uab.cat)
Institut d'Estudis del Treball (http://iet.uab.cat/)
Departament de Sociologia. Universitat Autònoma de Barcelona
pedro.lopez.roldan@uab.cat

Sandra Fachelli
Departament de Sociologia i Anàlisi de les Organitzacions
Universitat de Barcelona
Grup de Recerca en Eduació i Treball (http://grupsderecerca.uab.cat/gret)
Departament de Sociologia. Universitat Autònoma de Barcelona
sandra.fachelli@ub.edu

Edición digital: http://ddd.uab.cat/record/129382

1ª edición, febrero de 2015

Edifici B · Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallés) · Barcelona · España
Tel. +34 93 581 1676

Índice general

Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

PARTE III. ANÁLISIS

Capítulo III.3
Análisis descriptivo de datos con una variable

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Cómo citar este capítulo:

López-Roldán, P.; Fachelli, S. (2015). Análisis descriptivo de datos con una variable. En P. López-Roldán y S. Fachelli, *Metodología de la Investigación Social Cuantitativa*. Bellaterra. (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. Capítulo III.3. Edición digital: http://ddd.uab.cat/record/163559

Capítulo acabado de redactar en diciembre de 2015

Contenido

PARTE III
Capítulo 3

Análisis descriptivo de datos con una variable

Los datos cuantitativos producidos en un proceso de investigación o bien utilizados a partir de una fuente de datos secundaria se empiezan analizar siempre a partir de técnicas de análisis de datos más sencillas de tipo univariable, es decir, donde tratamos específicamente cada una de las variables independientemente, sin relacionarlas con las demás, y de donde podemos extraer las conclusiones substantivas básicas más relevantes de la información que se estudia. La lectura univariable de la información proporciona pues unos primeros resultados descriptivos del fenómeno estudiado. Pero además el análisis de una sola variable nos permitirá observar y comprobar otras características de nuestros datos, como la comprobación de la correcta identificación de los mismos o la exploración del cumplimiento de determinados supuestos o condiciones necesarias que son requeridas para realizar un adecuado análisis estadístico de los datos con diferentes técnicas, tanto univariables como bivariables y multivariables. En consecuencia, no por ser una primera fase del proceso de análisis es menos relevante, aunque sin duda la riqueza del análisis de relaciones entre variables para cuenta de nuestras hipótesis más elaboradas atraerá sobre todo el interés del investigador/a.

La lógica del análisis y de la investigación es la lógica de la comparación. Cuando comparamos podemos valorar y relativizar, dar significación. Para ello necesitamos un referente que actúe de contraste o de medida para la comparación. Puedo comparar cuántas personas tienen ingresos altos y cuántas bajos, los ingresos de las personas los puedo comparar con un valor dado: con el salario mínimo interprofesional o con los ingresos medios de toda la población, o puedo comparar los ingresos de grupos sociodemográficos diferentes: los varones en relación a las mujeres o los jóvenes en relación a los adultos, o puedo comparar entres sociedades distintas: el área metropolitana de Madrid y el de Barcelona o entre España y Argentina, o entre momentos distintos en el tiempo: el año actual con respecto al año anterior o antes y después de la crisis, o bien comparo los resultados estimados de ingresos entre dos

investigaciones. Veremos en este capítulo como el tratamiento de los datos desde el punto de vista univariable implica constantes ejercicios de comparación[1].

El punto de partida para el análisis de datos estadísticos es la matriz de datos, la matriz de datos original de unidades por variables. La información que contiene la matriz de datos requiere que se procese para resumir la ingente cantidad de información que contiene. Los datos del estudio del CIS de la matriz CIS3041 que utilizamos en este manual, inicialmente contiene 2.480 individuos y 210 variables, es decir, un total de 520.800 datos dispuestos cada uno en la "celda" que cruza cada individuo con cada variables (x_{ij}). Sin la ayuda del ordenador y de las técnicas de análisis los procesen esos datos no son todavía información. El análisis estadístico de los datos tiene como uno de sus objetivos precisamente reducir esa complejidad agrupando o resumiendo los datos a través de tablas, gráficos y medidas de resumen que nos ayudarán a analizar la información que contienen. Realizaremos básicamente dos tipos de ejercicios u operaciones de comparación: la organización y ordenación de los datos mediante tablas de distribución de frecuencias, que se pueden presentar también de forma gráfica, y mediante diversos cálculos aritméticos (suma, resta, multiplicación o división) que resumen e indican aspectos relevantes y sintéticos para describir el contenido de los datos.

En esta tarea será fundamental deberemos tener presente la escala o el nivel de medición de las variables, pues las propiedades de cada escala determinan qué operaciones y cálculos son posibles. Ya vimos en el capítulo II.1 que la aritmética sólo es posible aplicarla a las variables cuantitativas, por tanto, las técnicas de análisis que exigen esa propiedad se podrán aplicar solamente a las variables que lo permitan. Será posible por ejemplo calcular la media del variable de ingresos, porque tiene sentido hacerlo en una variable cuantitativa como esta, y no lo será en el caso de la variable sexo porque carece de sentido pensar cuál es la media entre varones y mujeres cuando conceptualmente no se entiende y cuando formalmente es una variable cualitativa sin propiedad de la distancia ni unidad de medida.

En este capítulo daremos cuenta de la parte descriptiva de un análisis de datos estadísticos, ahora univariable. En Estadística existe dos vertientes: la descriptiva y la inferencial. Ambas confluyen cuando trabajamos con datos de una muestra estadística pues la cuestión que se plantea es fundamental: ¿en qué medida las conclusiones que extraigo de las comparaciones y descripciones realizadas con los datos de la muestra se pueden extrapolar al conjunto de la población? La respuesta la ofrece la estadística inferencial.

Para ilustrar esta distinción, y a efectos comparativos, tomemos como ejemplo dos sondeos electorales que con motivo de la Elecciones Generales en España, celebradas el 20 de diciembre de 2015, estimaban la intención de voto de los ciudadanos en esos comicios. Completaremos el análisis comparado con los resultados electorales finalmente obtenidos[2]. En un caso presentamos los datos obtenidos de la encuesta realizada por el Centro de Investigaciones Sociológicas (CIS, estudio 3117) y en otro los obtenidos por Metroscopia en un sondeo para el diario EL PAIS. En ambas

[1] Este punto de vista es destacado en el texto de García Ferrando (1994: 45-64).
[2] Datos del Ministerio del Interior: http://www.interior.gob.es/informacion-electoral.

encuestas se seleccionó una muestra de ciudadanos elegida aleatoriamente entre la población mayor de 18 años con derecho a voto, a los que se les preguntó, entre otras cuestiones, sobre el posible sentido de su voto. En el Barómetro del CIS se encuestaron 17.452 personas mediante entrevista personal[3] y en la encuesta de Metroscopia a 2.800 personas a través de entrevista telefónica[4]. Los resultados obtenidos se presentan en la **¡Error! No se encuentra el origen de la referencia.**.

Tabla III.3.1. Distribución de frecuencias de la variable intención de voto y resultados de las Elecciones Generales de España de 2015

Candidatura	Sondeo Metroscopia		Barómetro del CIS		Resultados electorales		
	(1)	(2)	(1)	(2)	Votos	%	Escaños
PP	17,0	26,2	16,2	24,5	7.215.752	28,7	123
PSOE	12,8	19,8	14,9	22,5	5.530.779	22,0	90
Podemos(i)	13,6	21,0	11,8	17,8	4.781.093	19,0	69
Ciudadanos	11,8	18,2	11,6	17,5	3.500.541	13,9	40
UP/IU(ii)	3,2	4,9	2,6	3,9	923.133	3,7	2
Otros+Blanco(iii)	6,4	9,9	6,3	9,5	2.763.782	11,0	26
Total decidido	64,8	100,0	63,4	100,0	25.123.450	100,0	350
Indecisos	35,2	-	36,6	-	-	-	-
% Total	100,0	-	100,0	-	-	-	-
Total de casos	2.800	1.814	17.452	11.553	-	-	-

(1) Porcentaje de votos sobre el total de encuestados.
(2) Porcentaje de votos sobre el total de encuestados con voto decidido.
(i) Incluye a *Podemos, Compromís-Podemos-És el Moment, En Comú Podem* y *En Marea.*
(ii) *Unidad Popular: Izquierda Unida, Unidad Popular en Común*
(iii) Incluye, además de los votos en blanco, a *Esquerra Republicana de Catalunya-Catalunya Sí* (9 escaños), *Democràcia i Llibertat-Convergència-Demòcrates-Reagrupament* (8 escaños), *Euzko Alderdi Jeltzalea-Partido Nacionalista Vasco* (6 escaños), *Euskal Herria Bildu* (2 escaños), *Coalición Canaria-Partido Nacionalista Canario* (1 escaño).

Para hacer comparables los resultados de los sondeos y de las elecciones debemos tener presente que los datos de las encuestas presentan los porcentajes de los que declararon una intención de voto por una candidatura, en total suman el 64,8% y el 63,4% de los casos según cada encuesta, junto a los indecisos (columna (1) de la tabla), es decir, que nada más y nada menos que de alrededor de un 35% desconocemos su posible comportamiento electoral. Si recalculamos la distribución de porcentajes de los

[3] La información del avance del Barómetro se publica en http://www.cis.es/cis/opencms/ES/9_Prensa/ y los datos se publican en http://www.cis.es/cis/opencm/ES/11_barometros/index.jsp.
[4] Sondeo realizado por Metroscopia para el diaro EL PAIS y publicado el 14 de diciembre de 2015.
http://elpais.com/elpais/2015/12/11/media/1449862752_322272.html

FICHA TÉCNICA
Sondeo efectuado mediante entrevistas telefónicas a una muestra nacional de personas **mayores de 18 años.** Se han completado **2.800 entrevistas**, estratificadas por la intersección hábitat /comunidad autónoma y distribuidas de manera proporcional al total de cada región, con cuotas de sexo y edad aplicadas a la unidad última (persona entrevistada). Partiendo de los criterios del muestreo aleatorio simple, para un nivel de confianza del 95,5% (que es el habitualmente adoptado) y en la hipótesis más desfavorable de máxima indeterminación (p=q=50), el **margen de error de los datos referidos al total de la muestra es de ± 1,9 puntos.**
Recogida y tratamiento de la información realizada en Metroscopia. Realización del trabajo de campo: **del 7 al 10 de diciembre de 2015.**
Para la estimación del reparto de escaños, se han tomado adicionalmente en cuenta las 18.800 entrevistas llevadas a cabo separadamente por Metroscopia, durante los meses de noviembre y diciembre, **en 35 provincias** sobre muestras representativas de **400 a 800 electores**, según los casos. Las estimaciones de resultado electoral no constituyen una predicción de lo que los electores finalmente votarán, sino un intento de traducción, en términos de intenciones de voto (y su correlativa traducción en escaños), del estado de ánimo y de la predisposición que declaran en el concreto momento de la entrevista. Las estimaciones se basan en los datos directos obtenidos (intención directa de voto y simpatía o afinidad por algún partido entre quienes no la declaran) que se adjuntan. A partir de los mismos, y con criterios interpretativos distintos a los de Metroscopia, pueden alcanzarse estimaciones no plenamente coincidentes con las aquí ofrecidas.

partidos tomando como 100% el total de casos que manifiestan su intención de voto se obtienen los datos de la columna (2) de la tabla[5].

En ambos sondeos se obtienen una misma imagen general y acertada del cambio político devenido en España, con ciertas divergencias si comparamos los sondeos entre sí y con los resultados finales. En todos los casos se constata una distribución bastante repartida entre las distintas opciones políticas, si lo comparamos con los resultados electorales del pasado en España caracterizados por un modelo marcadamente bipartidista entre PSOE y PP. En las elecciones de 2011 ambas formaciones acumularon el 73,4% mientras que en 2015 representan el 50,7% de los votos.

Si comparamos los distintos partidos políticos entre sí observamos cómo el partido con mayor intención de voto era el PP, con el 26,2% y el 24,5% de personas que lo eligieron en cada una de las encuestas. Es un resultado relativamente certero: acertado por estimar correctamente el partido mayoritario pero en ambos casos infraestimando el resultado electoral que fue del 28,7%. El resto de formaciones aparecen distanciadas, de forma clara UP/IU, y con resultados más similares el resto, más de lo que luego fueron los resultados electorales, al sobrestimar el resultado de Ciudadanos en ambas encuestas en 4 puntos. El resultado esperado del PSOE y Podemos fue acertado, dentro de los márgenes de error que se esperaría en un estudio por muestreo estadístico.

Detengámonos un momento en los resultados de Metroscopia de la primera columna para explicar las implicaciones del margen de error. Tras el PP las tres candidaturas siguientes aparecen muy igualadas en intención directa de voto: entre el 11,8% de Ciudadanos, el 12,8% del PSOE y el 13,6% de Podemos. A partir de estos resultados nos preguntamos: ¿Se puede concluir de estas estimaciones que los resultados electorales darán como resultado este orden de partidos? ¿Se puede garantizar cuál será el partido más votado en segundo lugar? La respuesta es no. Independientemente del conocimiento de los resultados electorales, cuando los resultados son tan ajustados las diferencias estadísticas son muy pequeñas y entran dentro del margen de error que tiene todo muestreo estadístico. En estas condiciones se dice que estamos ante un empate técnico. La técnica, es decir, el muestreo estadístico y el sondeo realizado, no nos permite afirmar más que comparando los resultados éstos serán muy parecidos entre los tres partidos. Si nos fijamos en la ficha técnica del estudio podemos ver que el margen de error muestral es del 1,9%, un valor bajo, es decir, es un valor que nos determina a partir de qué diferencias puedo afirmar con garantías mínimas que un partido político obtendrá resultados por delante o por detrás de otros[6]. Las diferencias entre las tres formaciones son de un máximo del 1,8%, por debajo del margen permitido del 1,9%.

De aquí se concluye que a pesar de observar descriptivamente diferencias entre las candidaturas (los diferentes porcentajes), técnicamente, dado el error que implican los

[5] Este recálculo es legítimo pero no deja ser lo que se denomina una "imputación" de resultados: establecemos que ese 35% de indecisos tendrá un comportamiento idéntico al 65% que se pronunció. Y no necesariamente es así. Pero no existe otra alternativa con estos datos de intención directa para realizar la comparación. Una alternativa posible más elaborada es la aplicación de un determinado modelo de ajuste de los datos, la llamada "cocina", que en función de variables como el recuerdo de voto y la extrapolación en términos de escaños se efectúan cálculos más sofisticados.

[6] En el caso del Barómetro del CIS este error de muestreo es del 0,76%, inferior pues el tamaño de la muestra es notablemente superior.

datos que se obtienen de una muestra de la población, no tenemos garantías suficientes como para afirmar qué formación política quedará segunda, tercera y cuarta, cuando queremos inferir los resultados de la muestra al conjunto de toda la población electoral.

No obstante los resultados finales de las elecciones muestras más diferencias entre los partidos políticos de lo que vaticinaron las encuestas y de lo que implica el margen de error. Como hemos destacado en capítulos precedentes al hablar de la encuesta y el muestreo, el error total es la suma del error estadístico (el error muestral cuantitativo que hemos comentado) más el error sistemático, es decir, toda suerte de motivos por los cuales una determinada medición difiere del valor cierto poblacional: la alta proporción de indecisos, la incertidumbre del cambio político, características relacionadas con el trabajo de campo, etc. inciden igualmente en la precisión de los resultados estimados. De la comparación entre las estimaciones y los resultados del 20-D se concluye que parte de los votantes de Ciudadanos que se inclinaban por esta opción en la campaña electoral acabaron votando al PP y que una parte mayor de los indecisos también se inclinó en este sentido. Y este tipo de fenómenos no hay estudio que lo pueda determinar con precisión.

La conclusión, por ello, debe expresarse en sentido contrario. No es tanto cuánto nos hemos equivocado, que tampoco fue mucho, sino cuánto hemos acertado en la previsión dados los instrumentos a nuestro alcance. No le podemos pedir más a una encuesta de lo que puede dar y dio muchísimo.

En este ejemplo hemos intentado mostrar los dos aspectos del análisis de datos de una sola variable, en su vertiente descriptiva y en su vertiente inferencial, basando continuadamente nuestras lecturas e interpretaciones en la lógica de la comparación. Sobre los aspectos descriptivos profundizaremos en este capítulo y sobre la inferencia estadística en el próximo.

El objetivo de un análisis estadístico descriptivo es resumir, de la manera más concisa, comprensible y visual posible la información relevante que contienen los datos. Presentaremos en primer lugar la organización de los datos estadísticos de la matriz mediante la tabulación de los datos generando e interpretando tablas de distribución de frecuencia, para a continuación visualizar esta información con las representaciones gráficas. Por otro lado se presentará el cálculo y la interpretación de los indicadores descriptivos o medidas de resumen de la distribución de las variables, según la escala de medición. Dedicaremos un apartado específico a la exploración de los datos para complementar la descripción con otras presentaciones y cálculos de interés. Finalmente concluiremos este capítulo con algunos aspectos complementarios de interés en el tratamiento y análisis de la información de una variable.

1. Distribuciones de frecuencias

El primer instrumento para resumir un conjunto de datos de una variable es enumerar los diferentes valores observados de la misma y calcular la frecuencia o número de repeticiones que resulta del recuento de cada valor. La tabla que lista de forma ordenada los valores observados junto con las correspondientes frecuencias constituyen la distribución de frecuencias.

En la Tabla III.3.2 podemos ver la distribución de la variable nominal P37 (el estado civil de la persona entrevistada) y en la Tabla III.3.3 la de la variable ordinal ESTUDIOS (Estudios de la persona entrevistada), resultados extraídos con el software SPSS de la matriz de datos CIS3041.sav. En el primer caso vemos que la variable tiene 5 valores (categorías o modalidades) que se han codificado con los valores 1 a 5, junto con una etiqueta identificativa que precisa el estado civil asociado a cada valor. En la columna de frecuencia aparece el recuento del número de personas que está en cada situación. Como se puede apreciar tenemos 3 casos de valores perdidos por falta de información que corresponden a "no contesta". Como estos casos no se consideran válidos para el análisis estadístico tiene un doble tratamiento en la tabla de frecuencias, primero se incluyen para ver su importancia y luego se excluyen para realizar el análisis. Junto a la distribución de frecuencia absolutas tenemos la distribución de porcentajes, primero con todos los valores y luego con los valores válidos, eliminando los perdidos. Una última columna presenta los porcentajes acumulados. La tabla nos muestra que sobre las 2.477 personas entrevistadas de las que conocemos su estado civil poco más de la mitad están casadas (el 53,7%) y el 30,6% están solteras, mientras que el resto de situaciones, viudo/a, separado/a y divorciado/a, son minoritarias en comparación a aquellas dos más frecuentes.

Tabla III.3.2. Distribución de frecuencias de la variable P37
Estado civil de la persona entrevistada

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Casado/a	1331	53,7	53,7	53,7
	2 Soltero/a	758	30,6	30,6	84,3
	3 Viudo/a	198	8,0	8,0	92,3
	4 Separado/a	72	2,9	2,9	95,2
	5 Divorciado/a	118	4,8	4,8	100,0
	Total	2477	99,9	100,0	
Perdidos	9 N.C.	3	0,1		
	Total	2480	100,0		

Fuente: CIS, Estudio 3041

Tabla III.3.3. Distribución de frecuencias de la variable ESTUDIOS
Estudios de la persona entrevistada

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	1 Sin estudios	139	5,6	5,6	5,6
	2 Primaria	503	20,3	20,3	25,9
	3 Secundaria 1ª etapa	617	24,9	24,9	50,8
	4 Secundaria 2ª etapa	338	13,6	13,6	64,4
	5 F.P.	421	17,0	17,0	81,4
	6 Superiores	460	18,5	18,6	100,0
	Total	2478	99,9	100,0	
Perdidos	9 N.C.	2	0,1		
Total		2480	100,0		

Fuente: CIS, Estudio 3041

En el caso de la distribución de frecuencias de la variable de estudios vemos el porcentaje acumulado con los tres primeros niveles educativos alcanza a la mitad de los casos, el 50,8%, hasta tener estudios de secundaria de 1ª etapa, el nivel educativo más frecuente también. La otra mitad de la población tiene un nivel superior.

Las variables cualitativas en general contienen un número relativamente reducido de valores y requieren de una etiqueta identificativa. En el caso de las variables cuantitativas los valores hablan por sí mismos y no la requieren, pues tienen una unidad de medida, y suelen tener un conjunto numeroso de valores. Es el caso de la variable P32 (la edad) que se presenta en la Tabla III.3.4 a doble columna.

Tabla III.3.4. Distribución de frecuencias de la variable P32

Edad de la persona entrevistada

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado			Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	18	32	1,3	1,3	1,3	Válido	56	36	1,5	1,5	67,4
	19	32	1,3	1,3	2,6		57	42	1,7	1,7	69,1
	20	18	0,7	0,7	3,3		58	29	1,2	1,2	70,3
	21	28	1,1	1,1	4,4		59	27	1,1	1,1	71,4
	22	38	1,5	1,5	6,0		60	48	1,9	1,9	73,3
	23	25	1,0	1,0	7,0		61	27	1,1	1,1	74,4
	24	27	1,1	1,1	8,1		62	33	1,3	1,3	75,7
	25	51	2,1	2,1	10,1		63	39	1,6	1,6	77,3
	26	42	1,7	1,7	11,8		64	40	1,6	1,6	78,9
	27	40	1,6	1,6	13,4		65	48	1,9	1,9	80,8
	28	23	0,9	0,9	14,4		66	37	1,5	1,5	82,3
	29	39	1,6	1,6	15,9		67	39	1,6	1,6	83,9
	30	46	1,9	1,9	17,8		68	24	1,0	1,0	84,9
	31	48	1,9	1,9	19,7		69	31	1,3	1,3	86,1
	32	41	1,7	1,7	21,4		70	36	1,5	1,5	87,6
	33	47	1,9	1,9	23,3		71	27	1,1	1,1	88,7
	34	53	2,1	2,1	25,4		72	28	1,1	1,1	89,8
	35	51	2,1	2,1	27,5		73	18	0,7	0,7	90,5
	36	37	1,5	1,5	29,0		74	21	0,8	0,8	91,4
	37	48	1,9	1,9	30,9		75	19	0,8	0,8	92,1
	38	47	1,9	1,9	32,8		76	20	0,8	0,8	92,9
	39	46	1,9	1,9	34,6		77	18	0,7	0,7	93,7
	40	48	1,9	1,9	36,6		78	25	1,0	1,0	94,7
	41	43	1,7	1,7	38,3		79	16	0,6	0,6	95,3
	42	57	2,3	2,3	40,6		80	17	0,7	0,7	96,0
	43	61	2,5	2,5	43,1		81	17	0,7	0,7	96,7
	44	71	2,9	2,9	45,9		82	14	0,6	0,6	97,3
	45	51	2,1	2,1	48,0		83	15	0,6	0,6	97,9
	46	51	2,1	2,1	50,0		84	13	0,5	0,5	98,4
	47	45	1,8	1,8	51,9		85	11	0,4	0,4	98,8
	48	42	1,7	1,7	53,5		86	8	0,3	0,3	99,2
	49	45	1,8	1,8	55,4		87	4	0,2	0,2	99,3
	50	57	2,3	2,3	57,7		88	5	0,2	0,2	99,5
	51	33	1,3	1,3	59,0		89	4	0,2	0,2	99,7
	52	34	1,4	1,4	60,4		90	2	0,1	0,1	99,8
	53	49	2,0	2,0	62,3		91	2	0,1	0,1	99,8
	54	56	2,3	2,3	64,6		92	1	0,0	0,0	99,9
	55	34	1,4	1,4	66,0		94	3	0,1	0,1	100,0
							Total	2480	100,0	100,0	67,4

Fuente: CIS, Estudio 3041

Como se puede observar la distribución con un elevado número de valores, desde 18 hasta 94, motiva sea poco informativa, a diferencia de las variables cualitativas anteriores, y que necesitemos algún procedimiento para reducir más la información. En este caso adquiere mayor relevancia la columna del porcentaje acumulado que nos permite decir por ejemplo que la mitad de los entrevistados/as tiene hasta 46 años o que los jóvenes entre 18 y 25 años son el 10,1%.

Las tablas de frecuencias son la forma de presentar de forma resumidas la información de la distribución de una variable que puede tener cualquier escala de medición. En las tablas podemos incluir frecuencias absolutas (el recuento), relativas (en tanto por uno o proporciones) y en porcentaje (en tanto por cien). Veamos a continuación la notación que emplearemos para expresar la información de una tabla de frecuencias.

1.1. Notación de las tablas de frecuencias

Para expresar la información de una tabla de frecuencias seguiremos la siguiente notación que se ilustra en la Tabla III.3.5:

x_i A los distintos valores de la tabla los identificaremos por x_i donde x sería el nombre de la variable y el subíndice i nos indica de forma genérica que los valores (categoría, modalidad o clase) van desde el primer valor, $i=1$, hasta el último, $i=k$, esto es, que están indexados por $i=1,2,\ldots,k$, siendo k por tanto el número total de valores de la variable. Cuando dispongamos de datos agrupados en intervalos al valor que representa al intervalo se le denomina marca de clase o punto medio del intervalo.

k Es el número de valores distintos de la variable.

n_i Es la frecuencia absoluta del valor x_i, el recuento del número de veces que aparece el valor *i-ésimo*. El conjunto de estos valores constituye la distribución de frecuencias absolutas.

n Es el número total de casos (unidades, observaciones o individuos). Se cumple que la suma de las frecuencias absolutas de todos los valores, desde $i=1$ hasta $i=k$, es el número de casos y que expresamos así: $\sum_{i=1}^{k} n_i = n$.

A partir de las frecuencias absolutas se calculan las relativas. La distribución de frecuencias relativas es el resultado de dividir la frecuencia de cada valor por el número total de casos (la suma de frecuencias de todas las clases). Este cociente representa la proporción de casos que tienen un valor de clase sobre el total de casos:

f_i Frecuencia relativa o proporción: $f_i = \frac{n_i}{n}$.

Se cumple que la suma de las frecuencias relativas es 1: $\sum_{i=1}^{k} f_i = 1$.

La frecuencia relativa se expresa en tanto por uno. Pero habitualmente las frecuencias relativas no se presentan en proporciones sino en porcentajes, multiplicando sencillamente la proporción por 100.

p_i Frecuencia relativa porcentual o porcentaje: $p_i = f_i \times 100 = \frac{n_i}{n} \times 100$

Se cumple que la suma de las frecuencias relativas es 100: $\sum_{i=1}^{k} p_i = 100$.

Tabla III.3.5. Distribución de frecuencias de la variable P1

Valoración de la situación económica general de España

	Índice i=1...5 (k=5)	Valor	Etiqueta	Frecuencia absoluta	Frecuencia absoluta acumulada	Frecuencia relativa	Frecuencia relativa acumulada	Porcentaje	Porcentaje válido	Porcentaje acumulado
	i	x_i		n_i	N_i	f_i	F_i	p^*_i	p_i	P_i
Válido	i=1	1	Muy buena	0	0	0,000	0,000	0,0	0,0	0,0
	i=2	2	Buena	32	32	0,013	0,013	1,3	1,3	1,3
	i=3	3	Regular	399	431	0,161	0,174	16,1	16,1	17,4
	i=4	4	Mala	990	1421	0,401	0,575	39,9	40,1	57,5
	i=5	5	Muy mala	1050	2471	0,425	1,000	42,3	42,5	100,0
			Total	**2471**		**1,000**		**99,6**	**100,0**	
Perdidos		8	N.S.	5				0,2		
		9	N.C.	4				0,2		
			Total	9				0,4		
Total				**2480**				**100,0**		

Fuente: CIS, Estudio 3041

El interés de la distribución de frecuencias relativas en porcentaje radica en la ventaja de disponer de un valor de referencia, el valor 100, que estandariza la distribución a su tamaño facilitando la comparación de la distribución con cualquier otra independientemente del número de casos, es decir, considerando que el número de casos es 100. El total de la distribución de frecuencias relativas es la suma de todas las frecuencias relativas, que será 1, si se emplean proporciones, y 100 si se expresan en porcentajes.

Los datos revelan el momento especialmente difícil de la economía española en 2014 resultado del período de profunda crisis económica iniciado en 2008 que lleva a expresar a los entrevistados una valoración muy negativa de la situación económica general. De hecho nadie de las 2471 personas que contestaron eligió que la situación era muy buena, mientras que el 82,6% la valoró como mala o muy mala.

Cuando se calculan las frecuencias relativas los cocientes suelen presentarse con valores decimales. Habitualmente estos decimales se limitan a uno cuando se manejan porcentajes, si bien se pueden ajustar al número deseado[7]. Cuando decimos que se

[7] En Ciencias Sociales, por lo general, no necesitamos trabajar con un gran número de decimales, en particular cuando se presentan tablas descriptivas con porcentajes, con uno es suficiente, a lo sumo dos dependiendo de algún un caso muy particular (un decimal en porcentaje es el equivalente a tres decimales en frecuencias relativas o tanto por uno). El nivel de imprecisión que siempre tienen los datos, de encuesta en particular, no justifica la obsesión que a veces se da poner poner más decimales en los datos presentados que lo único que hacen es complicar la

limitan a un decimal de hecho procedemos a efectuar el redondeo del porcentaje a un decimal.

Conviene recordar los criterios que se aplican en el redondeo. Las reglas del redondeo se aplican al decimal situado en la siguiente posición al número de decimales que se quiere tener. Por ejemplo, si tenemos un número con 2 decimales y lo que remos redondear a 1 decimal se aplicarán las reglas siguientes:

- Si el dígito es menor que 5 el anterior no se modifica: 3,14 se redondea a 3,1.
- Si el dígito es mayor o igual que 5 el anterior se incrementa en una unidad: 3,55 y 3,56 se redondean a 3,6.

Por último disponemos de la distribución de frecuencias acumuladas, que tienen interés cuando los valores o categorías de las variables están ordenados, es decir, que se puede aplicar fundamentalmente a distribuciones de variables medidas a partir del nivel ordinal. Esta distribución nos indica para cada categoría o valor de clase el número de casos (o bien el porcentaje de casos) que quedan por debajo del límite real superior de clase de dicha categoría, es decir, es la frecuencia desde el primer valor hasta dicha categoría incluida. Aunque también se podrían considerar frecuencias acumuladas para los valores ordenados inversamente, de mayor a menor. Seguiremos las siguientes notaciones:

N_i Frecuencia absoluta acumulada hasta el valor x_i de la variable, es decir, la frecuencia de los valores menores o iguales que el valor *i-ésimo*, y se calcula mediante: $N_i = n_1 + n_2 + \cdots + n_i = \sum_{j=1}^{i} n_j$.

La última frecuencia absoluta acumulada, la que corresponde al valor más alto, es la suma de todas las frecuencias: $N_k = n$.

F_i Frecuencia relativa acumulada hasta el valor f_i de la variable, es decir, la frecuencia relativa de los valores menores o iguales que el valor *i-ésimo*, y se calcula mediante: $F_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^{i} f_j$, o bien a través de: $F_i = \frac{N_i}{n}$ ya que la primera expresión acumula más errores de redondeo. Se cumple que la última frecuencia relativa acumulada, la que corresponde al valor más alto, es la suma de todas las frecuencias e igual a 1: $F_k = 1$.

P_i Frecuencia relativa acumulada hasta el valor p_i de la variable, es decir, la frecuencia relativa de los valores menores o iguales que el valor *i-ésimo*, y se calcula mediante: $P_i = F_i \times 100 = \frac{N_i}{n} \times 100$.

Se cumple que la última frecuencia relativa acumulada, la que corresponde al valor más alto, es la suma de todas las frecuencias e igual a 100: $P_k = 100$.

claridad de la información. Otra cuestión es cuando se aplican técnicas de análisis y pruebas estadísticas donde puede tener todo su sentido.

1.2. Agrupación de valores

Las distribuciones de frecuencias se pueden presentar directamente sin ninguna agrupación, pero cuando el número de valores de la variable es alto disponemos de una gran cantidad de información en términos de valores diferentes cuya extensión motiva que sea necesario simplificarla optando por una distribución de frecuencias agrupadas. Se trata de resumir y hacer más apreciable el patrón de comportamiento de las variables, de ganar en significación a costa de una pérdida de información. La agrupación de los valores de las variables puede aplicarse tanto a las variables cuantitativas como cualitativas. En este último caso las razones de la agrupación no suele ser la cantidad de información, aunque también se dispone de variables categóricas con muchos valores, sino criterios conceptuales de categorización o criterios empíricos de frecuencias. En el caso de las variables cuantitativas se convierte en una necesidad disponer de la información agrupada, en este caso a partir de la construcción de intervalos de valores y así disponer de una tabla de frecuencias, así como de gráficos denominados histogramas que veremos posteriormente, que facilitan la lectura de la información.

La Tabla III.3.6 muestra la estructura por edades de la población en España con datos agrupados en intervalos (en quinquenios) a partir de la información del Censo de Población del INE. Se comparan dos momentos en el tiempo, 1981 y 2011, y también entre varones y mujeres[8].

Tabla III.3.6. Distribución de frecuencias de la edad (en quinquenios) según el sexo. España. Censo de Población de 1981 y 2011

	Censo de 1981						**Censo de 2011**					
Edad	Varones	%	Mujeres	%	**Total**	%	Hombres	%	Mujeres	%	**Total**	%
De 0 a 4 años	1.583.910	8,6	1.491.443	7,8	**3.075.352**	**8,2**	1.280.210	5,5	1.199.985	5,1	**2.480.195**	**5,3**
De 5 a 9 años	1.703.919	9,2	1.604.131	8,4	**3.308.049**	**8,8**	1.229.078	5,3	1.160.767	4,9	**2.389.845**	**5,1**
De 10 a 14 años	1.695.477	9,2	1.606.851	8,4	**3.302.328**	**8,8**	1.131.354	4,9	1.067.978	4,5	**2.199.332**	**4,7**
De 15 a 19 años	1.665.836	9,0	1.597.476	8,3	**3.263.312**	**8,7**	1.134.415	4,9	1.068.586	4,5	**2.203.001**	**4,7**
De 20 a 24 años	1.480.485	8,0	1.461.693	7,6	**2.942.178**	**7,8**	1.278.158	5,5	1.237.288	5,2	**2.515.447**	**5,4**
De 25 a 29 años	1.278.895	6,9	1.258.533	6,6	**2.537.428**	**6,7**	1.560.396	6,8	1.527.211	6,4	**3.087.607**	**6,6**
De 30 a 34 años	1.230.896	6,7	1.224.419	6,4	**2.455.314**	**6,5**	1.995.991	8,6	1.897.737	8,0	**3.893.728**	**8,3**
De 35 a 39 años	1.126.499	6,1	1.119.307	5,8	**2.245.806**	**6,0**	2.107.543	9,1	1.981.569	8,4	**4.089.112**	**8,7**
De 40 a 44 años	1.017.661	5,5	1.038.349	5,4	**2.056.009**	**5,5**	1.963.891	8,5	1.878.394	7,9	**3.842.285**	**8,2**
De 45 a 49 años	1.167.417	6,3	1.193.807	6,2	**2.361.225**	**6,3**	1.817.822	7,9	1.787.114	7,5	**3.604.936**	**7,7**
De 50 a 54 años	1.109.111	6,0	1.155.980	6,0	**2.265.091**	**6,0**	1.605.711	6,9	1.618.434	6,8	**3.224.145**	**6,9**
De 55 a 59 años	985.136	5,3	1.052.866	5,5	**2.038.002**	**5,4**	1.334.513	5,8	1.372.010	5,8	**2.706.522**	**5,8**
De 60 a 64 años	722.572	3,9	873.971	4,6	**1.596.543**	**4,2**	1.194.961	5,2	1.268.449	5,3	**2.463.410**	**5,3**
De 65 a 69 años	632.122	3,4	813.485	4,2	**1.445.606**	**3,8**	1.035.437	4,5	1.141.745	4,8	**2.177.182**	**4,7**
De 70 a 74 años	511.004	2,8	702.804	3,7	**1.213.807**	**3,2**	790.452	3,4	932.328	3,9	**1.722.780**	**3,7**
De 75 a 79 años	335.782	1,8	516.398	2,7	**852.180**	**2,3**	760.258	3,3	999.223	4,2	**1.759.480**	**3,8**
De 80 a 84 años	163.455	0,9	298.505	1,6	**461.960**	**1,2**	522.078	2,3	801.583	3,4	**1.323.661**	**2,8**
85 y más años	81.564	0,4	181.606	0,9	**263.171**	**0,7**	263.761	1,1	505.489	2,1	**1.133.247**	**2,4**
Total	**18.491.741**	**100**	**19.191.622**	**100**	**37.683.363**	**100**	**23.104.303**	**100**	**23.711.613**	**100**	**46.815.916**	**100**

Fuente: Instituto Nacional de Estadística

[8] De hecho no se trata de un análisis univariable, la tabla es multidimensional o multivariable pues se consideran tres variables simultáneamente: la edad, el sexo y el tiempo.

Excepto el último intervalo, todos los demás tienen una amplitud de cinco años, aunque nos pueda parecer que tienen una amplitud de cuatro. Por ejemplo, el primer intervalo incluye todas las personas desde el momento de nacer hasta el momento justo antes de cumplir cinco años, tiene, por tanto, una amplitud de cinco años. El último intervalo está abierto o indefinido, y contiene todas las edades de 85 años y más, sin precisar el máximo.

De la lectura de los datos se desprende fácilmente que el porcentaje de varones en las primeras edades es superior, pues nacen más niños que niñas, e inversamente, la mayor proporción de mujeres en grupos etarios mayores, pues su esperanza de vida es mayor. De la comparación entre 1981 y 2011 se concluye el progresivo envejecimiento de la población española y los efectos de la reducción del número de nacimientos.

Los intervalos o clases tienen que estar ordenados y enganchados, un intervalo empieza donde acaba el anterior, y sean exhaustivos para contemplar todos los posibles valores de la variable.

Las agrupaciones de valores con variables cuantitativas implican construir intervalos de clase que podemos identificar con I_i. Consideraremos intervalos de igual amplitud cada uno de los cuales quedará determinado por dos valores: el límite inferior del intervalo (L_i) y el límite superior del intervalo (L_{i+1}). Los intervalos se suele escribir explicitando los límites entre un corchete y un paréntesis de la siguiente forma:

$$I_i = \left[L_i, L_{i+1}\right)$$

que se interpreta diciendo que el intervalo incluye los valores más grandes o iguales que L_i y estrictamente menores que L_{i+1}. Así la amplitud del intervalo será:

$$a_i = L_{i+1} - L_i$$

mientras que el punto medio del intervalo, llamado marca de clase, es:

$$x_i = \frac{L_{i+1} - L_i}{2}$$

Para ilustrar estos conceptos pongamos de ejemplo que disponemos de las notas de un grupo de alumnos de una asignatura. Las notas de cada alumno/a se pueden agrupar en intervalos iguales de amplitud 1, de la forma siguiente:

[0, 1), [1, 2), [2, 3), [3, 4), [4, 5), [5, 6), [6, 7), [7, 8), [8, 9), [9, 10].

Un alumno/a que obtiene un 8 en la asignatura se sitúa en el intervalo [8,9), igualmente si obtiene un 8,7 o un 8,9, pero si la nota es de 9 entonces se sitúa en el siguiente intervalo, el [9,10).

Las marcas de clase de cada intervalo serían:

0,5 1,5 2,5 3,5 4,5 5,5 6,5 7,5 8,5 9,5

Como suele ser habitual en nuestro entorno las notas se agrupan para convertirlas en evaluaciones "ordinales" constituyendo intervalos de diferente amplitud de la forma siguiente:

[0, 5) Suspenso
[5, 7) Aprobado
[7, 9) Notable
[9, 10) Sobresaliente o Matrícula de honor

El número de intervalos puede seguir criterios arbitrarios establecidos y justificados por el investigador/a. Suele ser habitual que los intervalos sean entre 5 y 20, dependiendo de los datos y del número de casos. No obstante también se han propuesto diversos métodos de cálculo:

a) Si n es el números de casos, y n es pequeño ($n \leq 100$), entonces el número de intervalos k se calcula aproximando al entero la raíz cuadrada de n: $k \approx \sqrt{n}$.

b) Si n es grande ($n > 100$), entonces se propone la fórmula de Sturgess: $k \approx \frac{\log(n)}{\log(2)}$

La amplitud de los intervalos, una vez determinado k, se calcula dividiendo el rango de valores de la variable por el número de intervalos:

$$a_i = \frac{x_{\max} - x_{\min}}{k}$$

Una vez determinada la amplitud se fijan los límites de todos los intervalos y se efectúa el recuento del número de casos de cada uno de ellos. El software estadístico resuelve de forma automática el problema de determinar el número de intervalos.

Hemos considerado intervalos de igual amplitud. Alternativamente se pueden construir intervalos de diferentes amplitudes. Pero en este caso es necesario corregir las frecuencias para garantizar que el área del rectángulo se corresponde con una frecuencia, con una altura, proporcional a la amplitud del intervalo intervalo.

Si el área del rectángulo es: $área = base \times altura = a_i \times h_i$, dado que la amplitud es variable, se trata de calcular para cada intervalo la altura h_i, su frecuencia, de forma proporcional a esa amplitud:

$$h_i = \frac{n_i}{a_i}$$

Obteniendo así la frecuencia por unidad de amplitud, de donde:

$$área = base \times altura = a_i \times h_i = a_i \times \frac{n_i}{a_i} = n_i$$

La frecuencia h_i se conoce también con el nombre de densidad de frecuencia.

Cuando se definen los intervalos es importante fijar claramente los límites de cada uno, garantizando que estén unidos formando un continuo. Esta idea nos lleva a hablar de correcciones por continuidad.

Si consideramos el ejemplo de la edad, medida en años, cuando efectuamos agrupaciones en intervalos como los siguientes:

Edad	Edad
0-4	[0,5)
5-17	[5,18)
18-64	[18,65)
65-109	[65,110)

Aparentemente no forman un continuo, cuando en realidad sí lo forman. De hecho, de 0 a 4 años no es una amplitud de 4 años, sino de 5, pues los niños/as que tienen 4 años comprenden los que acaban de cumplir 4 años y los que están a punto de cumplir 5, por tanto, la amplitud del intervalo es de 5, la de todos los niños/as hasta que cumplen 5 años (0 años, 1 año, 2 años, 3 años y 4 años). La forma más precisa de representar este intervalo aparente es mediante [0,5) que incluye los/as recién nacidos hasta los que se encuentran en el momento de cumplir 5 años. En el momento del cumpleaños pasan al siguiente intervalo [5-17)[9].

En otros casos en estos límites aparentes que conviene fijar de forma exacta, se aplica el criterio de extender el límite del intervalo 0,5 unidades a cada lado del intervalo. Un intervalo aparente, por ejemplo, de 10 a 15 tendría un intervalo exacto de [9,5-15,5).

2. Representaciones gráficas

La representación gráfica de la información estadística constituye una de las herramientas más atractivas para presentar de forma directa e inmediata los resultados de cualquier investigación. Existe una gran variedad de recursos gráficos que se han potenciado a partir del desarrollo de los programas informáticos y la capacidad gráfica de los ordenadores.

Los gráficos se pueden "editar" con el objetivo de hacerlos informativos y atractivos en relación a la características que emergen de las variables. Pero conviene tener presente hasta qué punto los gráficos son fiel reflejo de las conclusiones e interpretaciones de los resultados de un estudio o bien en nuestro afán de defender un argumento alteramos y desvirtuamos esa información.

Existe una amplia gama de representaciones gráficas. La opción que escojamos en cada momento tiene que adaptarse en todo caso al tipo de escala de medida de la variable implicada.

Cuando el gráfico representa distribuciones de frecuencias de las variables éstas contienen básicamente la misma información que la tabla de frecuencias pero con la ventaja muchas veces de poder visualizar más rápido y mejor la información de la distribución. Las representaciones gráficas de frecuencia tienen la particularidad de que

[9] Con el cambio de milenio se planteó una situación similar, había quien quería empezar el siglo XXI el 1-1-2000 cuando todavía no se había acabado el siglo XX, hubo que esperar hasta el 31-12-2000.

la forma del gráfico es la misma ya se consideren frecuencias absoluta, relativas i porcentajes, los único que cambia es la escala.

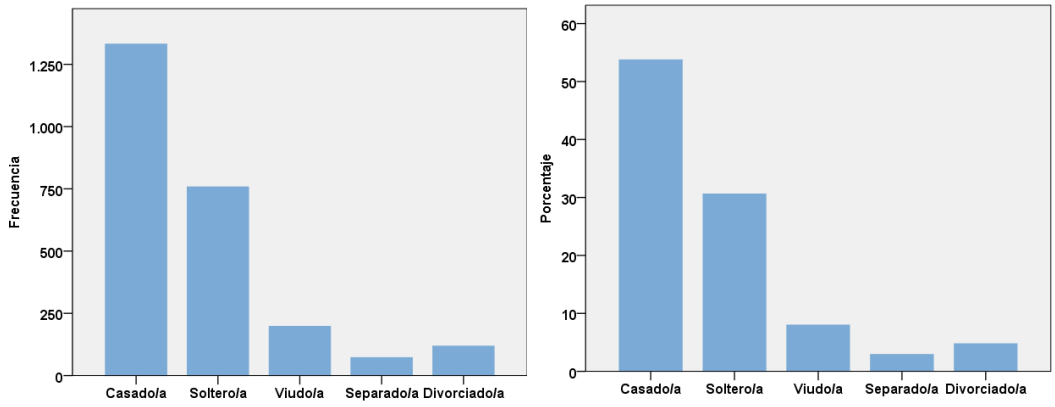
En los gráficos se considera convencionalmente que los valores de las variables se sitúen horizontalmente, en el llamado eje de abscisas, eje de categorías o eje *x*. Las frecuencias de los valores se sitúan verticalmente, en el eje de ordenadas o eje *y*. Además existe la posibilidad de intercambiarlos, de trasponerlos.

Presentaremos a continuación algunos de las representaciones más habituales según su adecuación a cada nivel de medición de las variables.

2.1. Gráficos para variables nominales

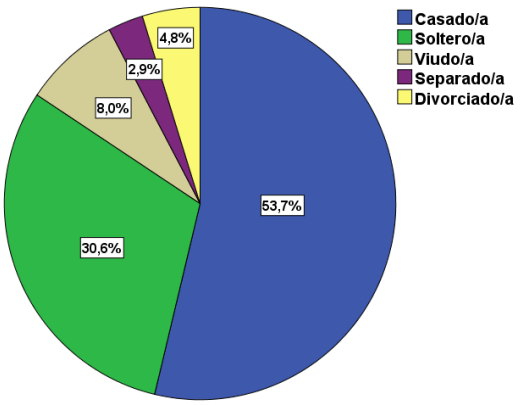
La representación gráficas más adecuada para una variable nominal es el gráfico o diagrama de barras. Se obtiene levantando para cada valor de la variable una barra de altura igual o proporcional a la frecuencia (absoluta, relativa o porcentaje). En el **¡Error! No se encuentra el origen de la referencia.** se representa el gráfico de barras de la variable P37 (Estado civil de la persona entrevistada) del estudio 3041 del CIS, primero con las frecuencias absolutas y a continuación con los porcentajes. Podemos ver cómo ambos gráficos son idénticos y tan solo cambian los valores de la escala: absolutos o porcentajes. En ambos casos la lectura es la misma, casados y solteros son las categorías más frecuentes de estado civil, mientras que las personas viudas, separadas o divorciadas son muchas menos.

Gráfico III.3.1. Gráfico de barras de la variables P37 (Estado civil de la persona entrevistada). Frecuencias absolutas y porcentajes



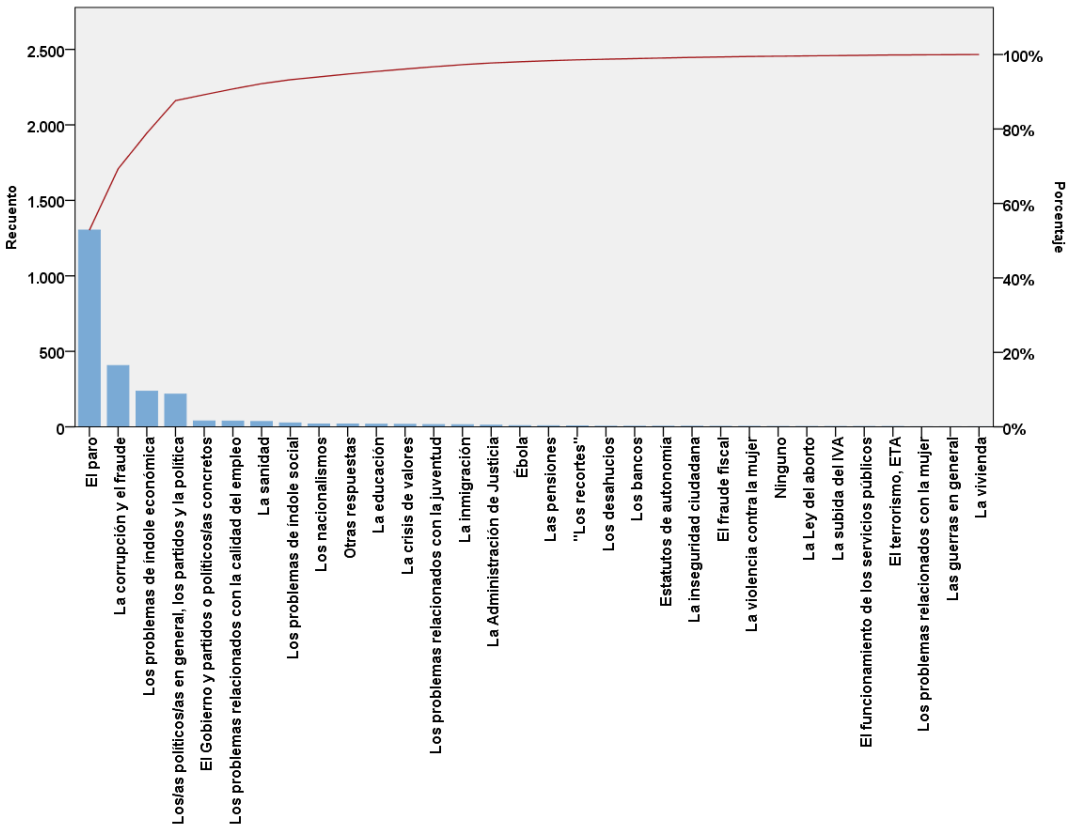
La misma información la podemos representar con gráficos de sectores (o circulares, a veces llamados también de pastel o torta) como aparece en el Gráfico III.3.2. Para obtener el gráfico se trata de repartir el área del círculo de 360º de forma proporcional a la frecuencia (absoluta, relativa o porcentual). A cada valor o categoría le corresponde un sector, determinado los grados con el siguiente cálculo: $grados = \frac{n_i}{n} \times 360^o$

Gráfico III.3.2. Gráfico de sectores de la variables P37 (Estado civil de la persona entrevistada). Porcentajes



El gráfico de Pareto es una representación en honor del economista y sociólogo Vilfredo Pareto (1848-1923) y del principio que lleva su nombre: *pocos vitales, muchos triviales*, que se puede expresar diciendo "el 80% de los problemas se resuelven con el 20% de les causas". Con el gráfico se trata de poner de manifiesto la concentración de la distribución de les frecuencias en unos pocos valores. La representación de hecho es un gráfico de barras que se ordenan según el orden de frecuencia los valores de la variable con un eje de frecuencias absolutas. En el gráfico se incorpora además una línea con los porcentajes acumulados con un segundo eje de porcentajes.

Gráfico III.3.3. Gráfico de Pareto de la variable P701 (Primer problema de España)

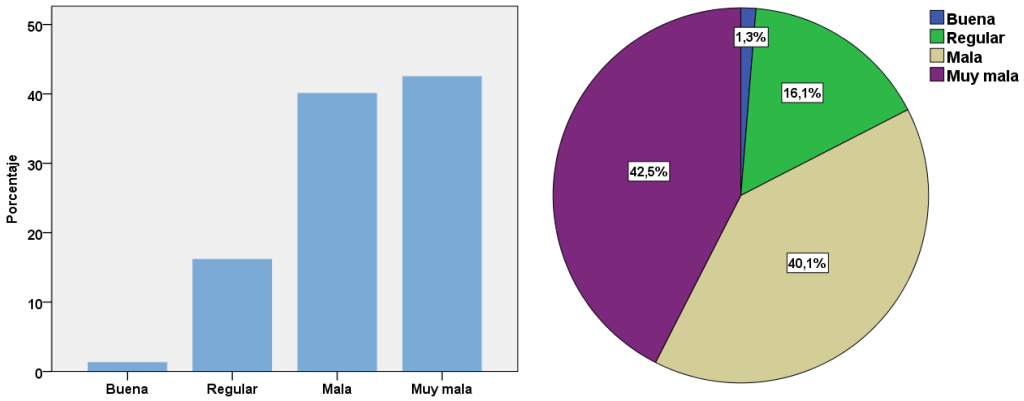


En el Gráfico III.3.3 se representa la variable P701 que expresa las principales preocupaciones de la ciudadanía a través de la pregunta de cuál es el principal problema que existe actualmente en España. A finales de 2014 los cuatro primeros problemas eran, por este orden: el paro, la corrupción y el fraude, los problemas de índole económica y los partidos y la política. Las cuatro acumulan prácticamente el 90% de las respuestas dadas por los entrevistados/as.

2.2. Gráficos para variables ordinales

Las variables nominales y ordinales comparten los mismos tipos de gráficos por ser variables cualitativas, categóricas, donde la propiedad de la distancia está ausente. Gráfico III.3.4 hemos representado el gráfico de barras y de sectores de la variable ordinal P1 (Valoración de la situación económica general de España), donde se puede observar, como vimos en la tabla de frecuencias, la valoración claramente negativa.

Gráfico III.3.4. Gráfico de barras y de sectores de la variable P1 (Valoración de la situación económica general de España). Porcentajes



Con las variables ordinales tiene sentido interpretar las frecuencias acumuladas y es posible obtener gráficos con esta información como el de ojiva que veremos a continuación con las variables cuantitativas.

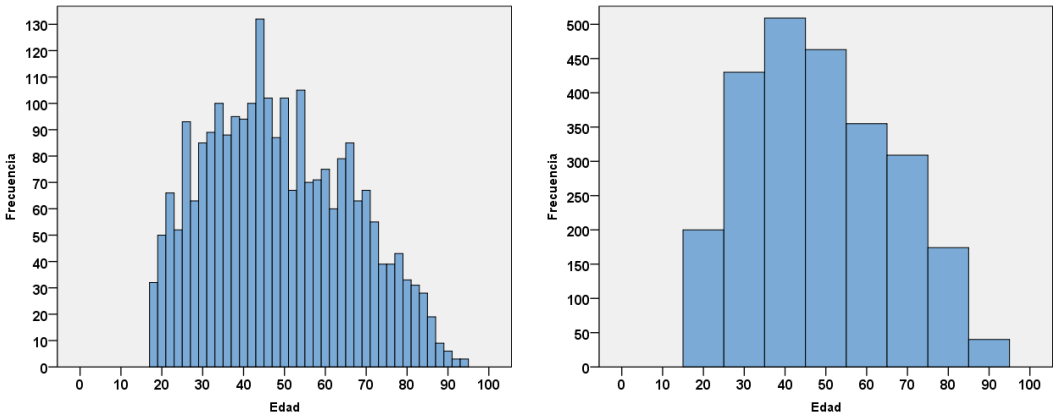
2.3. Gráficos para variables cuantitativas

Las representaciones gráficas que hemos visto hasta hora tienen en cuenta la existencia de una serie de categorías que expresan valores cualitativos sin una unidad de medida y por tanto sin una distancia entre ellos que se pueda cuantificar. Ahora las posibilidades gráficas se amplían por el hecho de ser cuantitativas aunque siempre tendremos la opción, si se considera adecuada, de representar alguna variable de escala numérica con algunos de los gráficos que hemos visto hasta ahora. No obstante, las variables cuantitativas que tienen un número importante de valores no se representan adecuadamente en los gráficos anteriores y se precisa agruparlos previamente. Los nuevos gráficos que veremos a continuación permiten esa agrupación a partir de la construcción de intervalos de valores, de forma automatizada cuando los realizamos con la computadora.

La representación gráfica habitual es el histograma que contiene la distribución de la variable a partir de levantar, para cada intervalo consecutivo que se construye, un rectángulo de área proporcional a la frecuencia. Estos rectángulos se disponen de forma adyacente y suelen ser de igual amplitud, determinando así su altura la presencia de más o menos casos en casos dentro del intervalo. El eje horizontal del gráfico dispone los intervalos mientras que el vertical se refiere a las frecuencias pero en términos de densidad: es el número de casos por unidad de la variable que se dispone en el eje horizontal.

En el Gráfico III.3.5 se representa el histograma de la variable edad con dos versiones, la propuesta inicialmente y de forma automatizada por el software estadístico, en este caso el SPSS, y una agrupación personalizada elaborada por nosotros. En el primer caso el ancho de intervalo es menor y en el segundo hemos construido intervalos de una anchura de 10 años, indicando que el intervalo empezara en los 15 años (si bien nuestros datos se inician en 18). En consecuencia la escala en el eje vertical será diferente en cada caso para reflejar la mayor o menor densidad de casos de cada rectángulo en cada histograma.

Gráfico III.3.5. Histograma de la variable P34 (la edad)



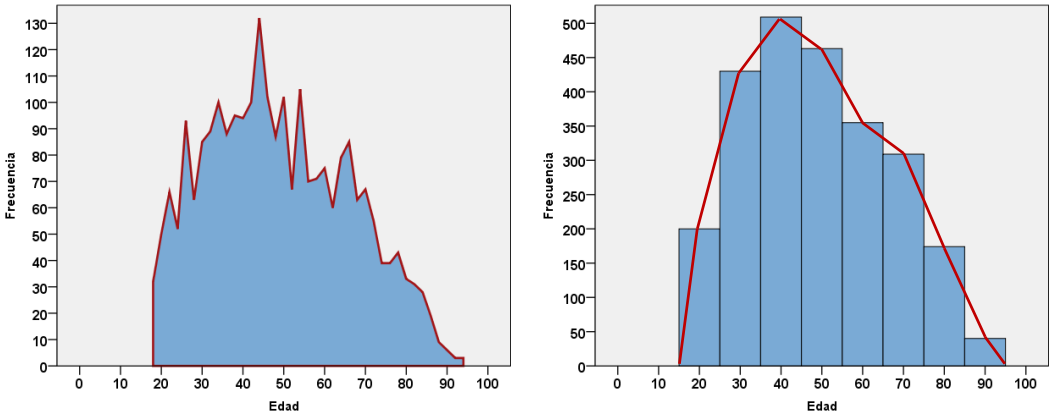
La forma y el contorno de los gráficos varía algo de un gráfico a otro pero en ambos casos el área total que comprenden todos los rectángulos es la misma, es una misma unidad resultado de sumarlos todos: la podemos expresar en términos de casos (suman 2800 y el área total sería de 2800), en términos de proporciones (suman 1 y el área es de 1) o en términos de porcentajes (suma 100 y el área total es 100).

Del gráfico se concluye claramente que a medida que aumenta la edad, en un primer momento y hasta más o menos los 50 años el número de personas va aumentando mientras que a partir de esa edad, por "ley de vida", el número de persona va progresivamente decreciendo. Se da además una cierta mayor concentración de casos en las edades más jóvenes hasta esos 50 años aproximadamente.

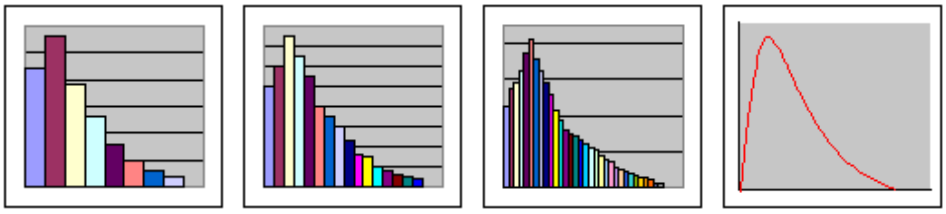
Una representación gráfica también de interés con variables cuantitativas es el polígono de frecuencias. Se construye a partir del histograma uniendo mediante una línea poligonal los puntos medios de las bases superiores de los rectángulos (sus marcas de clase). En el Gráfico III.3.6 se presenta de nuevo la distribución de la edad bajo la

forma de polígono de frecuencias, en el segundo caso, con las edad más agrupada, se presenta superpuesto con el histograma.

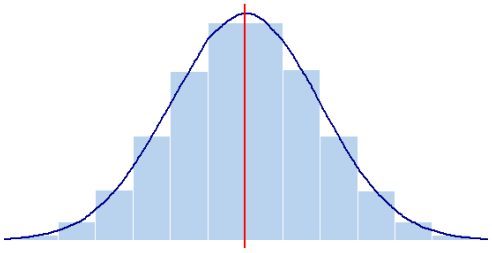
Gráfico III.3.6. Polígono de frecuencias de la variable P32 (edad)



Para una variable cuantitativa que incrementa el número de casos y el número de intervalos, haciéndolos cada vez de menor amplitud, el contorno del histograma tiende a suavizarse progresivamente hasta dibujar una línea continua que expresaría idealmente el perfil de la distribución.



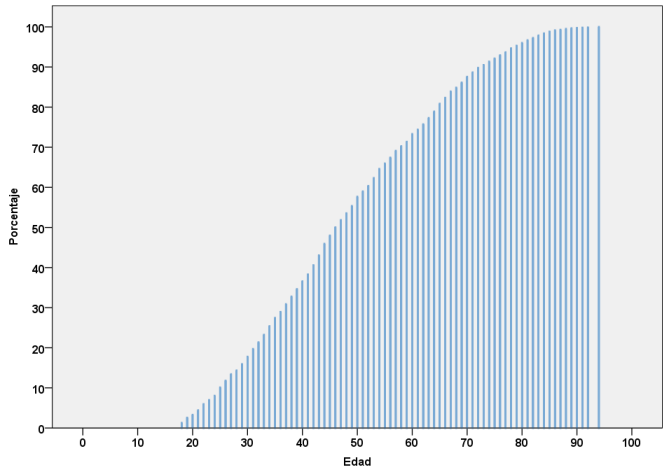
Representaremos en diversas ocasiones los histogramas con esa línea continua idealizando el carácter de la distribución. Veremos también que algunas de estas distribuciones ideales se corresponden con funciones matemáticas (funciones de densidad) que dibujan esa forma ideal continua del histograma. Será objeto de tratamiento sobre todo en el próximo capítulo al hablar de inferencia estadística y al tratar con distribuciones teóricas. Entre ellas destaca por su importancia en la teoría estadística la distribución normal cuya forma es perfectamente simétrica con forma de campana:



La ojiva es un polígono de frecuencias acumuladas, es decir, un diagrama con una línea poligonal siempre creciente. Con datos porcentuales la línea crece desde el 0% hasta el 100%. El Gráfico III.3.7 no se representa la línea pero se deriva a partir de un histograma de frecuencias acumuladas. En este tipo de gráfico se puede apreciar

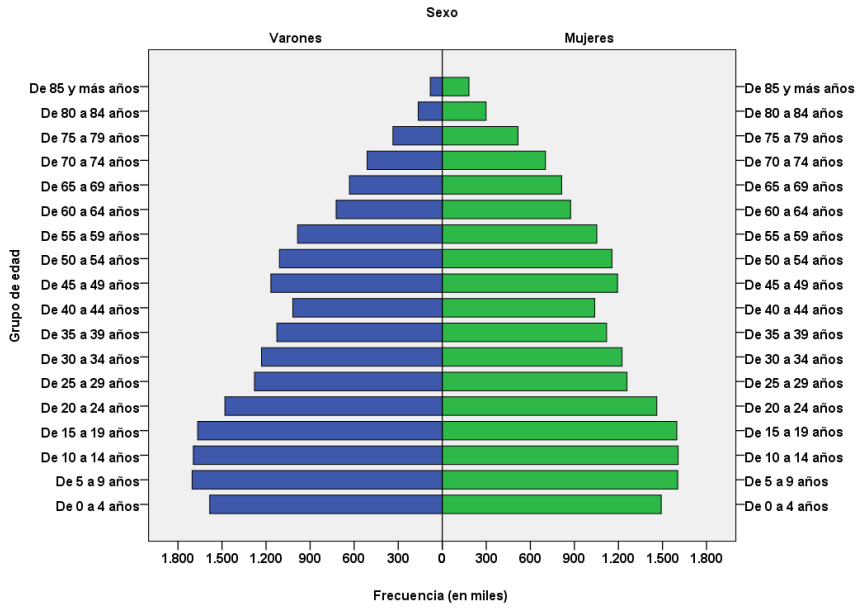
visualmente cómo, por ejemplo, el 50% de las personas entrevistadas se acumula hasta los 45 años de edad.

Gráfico III.3.7. Histograma de frecuencias acumuladas de la variable P32 (edad)



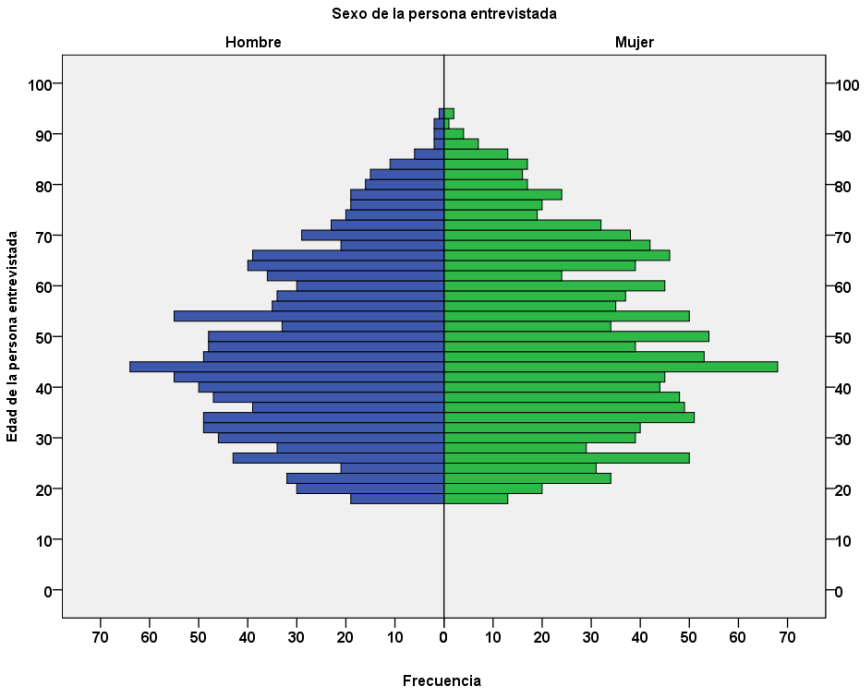
Los histogramas, y también los gráficos de barras, se pueden representar en forma de pirámides. Cuando la información que se maneja es la edad y la distribuimos entre varones y mujeres generamos una pirámide de población. El Gráfico III.3.8 presenta esta representación con los datos de la Tabla III.3.6 para el año 2011. En este caso se trata de un gráfico de barras en forma de pirámide.

Gráfico III.3.8. Pirámide de población. Distribución de la edad (en quinquenios) según el sexo. España. Censo de Población de 2011



El Gráfico III.3.9 reproduce la pirámide de población con los datos del estudio del CIS 3041 a partir de un histograma donde automáticamente se ha determinado la amplitud de los intervalos.

Gráfico III.3.9. Pirámide de población. Distribución de la edad (en quinquenios) según el sexo. España. Censo de Población de 2011



Cuando introduzcamos más adelante el análisis exploratorio de datos veremos otra representación de interés tan relevante como el histograma para dar cuenta de la distribución de una variable cuantitativa: el diagrama de caja. Para ello necesitamos introducir algunos conceptos previos relacionados con las medidas de resumen de las características de una distribución de frecuencias.

3. Características de una distribución de frecuencias

Las tablas de distribución de frecuencias junto a las representaciones gráficas constituyen una primera forma de dar cuenta de la información de las variables de una matriz de datos, ahora tratadas de forma univariable. Como decíamos al inicio, este es un primer tipo de ejercicio de comparación resultado de la organización y ordenación de los datos. Una segunda forma de realizar comparaciones es a través del cálculo de diversas medidas de resumen de las características de la distribución de una variable destinada a obtener información significativa en la descripción de las variables a través de un valor sintético. Esta operación consistirá en el cálculo aritmético de diferentes medidas, denominados estadísticos, que nos resumen los datos en unos pocos valores para dar cuenta de las características de las variables analizadas y de su distribución.

A continuación veremos que las características de una distribución univariable se podrán describir por medio de su posición (medidas de tendencia central y otras de posición), su variación (medidas de dispersión) y su forma. A continuación daremos cuenta de cada una de ellas. Pero es importante tener presente desde un inicio que estas medidas tiene sentido calcularlas dependiendo del nivel de medición de las variables. Según las propiedades de la escala de medición determinadas operaciones aritméticas no son pertinentes. Como podemos en la Tabla III.3.7, las variables medidas a nivel

cuantitativo, donde la propiedad de la distancia es definitoria, pueden calcularse las diferentes medidas de posición, dispersión y forma; sin embargo, en las variables cualitativas, muchas de estas medidas no son interpretables y no se calculan[10].

Tabla III.3.7. Medidas de resumen según la escala de medición

Medida de resumen		Escala de medición		
		Cualitativa		Cuantitativa
		Nominal	Ordinal	
Posición central	Moda	✓	✓	✓
	Mediana		✓	✓
	Media			✓
Posición	Valores extremos			✓
	Percentil		✓	✓
Dispersión	Rango			✓
	Rango intercuartil			✓
	Varianza			✓
	Desviación típica			✓
	Coeficiente de variación			✓
Forma	Asimetría			✓
	Curtosis			✓

Presentaremos a continuación estos estadísticos ejemplificándolos con las diversas variables que se recogen en la Tabla III.3.8.

Tabla III.3.8. Medidas de resumen de las variables P37, P701, P31, ESTUDIOS y P32 del Barómento del CIS y PIB per cápita del IDH de Naciones Unidas

Estadístico		P37 Estado civil	P701 Primer problema	P31 Sexo	P1 Valoración situación económica	ESTUDIOS Estudios	P32 Edad	PIB per cápita
n	Válidos	2477	2466	2480	2471	2478	2480	190
	Perdidos	3	14	0	9	2	0	5
Moda		1	1	2	5	3	44	443,96
Mediana		-	-	-	4	3	46	10338,97
Media		-	-	-	-	-	48,32	16486,51
Mínimo		-	-	-	-	-	18	443,96
Máximo		-	-	-	-	-	94	119029,12
Percentiles	25	-	-	-	4	2	34	3629,59
	50	-	-	-	4	3	46	10338,97
	75	-	-	-	5	5	62	22025,64
Rango		-	-	-	-	-	76	118585,16
Desviación típica		-	-	-	-	-	17,489	18383,00
Varianza		-	-	-	-	-	305,86	337934746,58
Asimetría		-	-	-	-	-	0,265	2,222
Curtosis		-	-	-	-	-	-0,790	6,677

[10] Una observación que el lector debería tener presente en su trabajo de análisis con datos estadísticos: el ordenador es capaz de efectuar cualquier cálculo, independientemente de la escala, aunque no tenga sentido.

3.1. Medidas de tendencia central

Las medidas de tendencia central o de posición central nos permiten caracterizar las distribuciones de las variables con la información de los valores que se sitúan en el "centro" de las mismas. La noción de centro se puede traducir en diferentes medidas por lo que existen diferentes características de tendencia central. Hablaremos de tres de ellas: la moda, la mediana y la media.

3.1.1. *La moda*

La moda (*Mo*) es el valor más frecuente de la variable. Es la modalidad "que más se lleva", la categoría que más veces se repite. Puede resultar que una distribución presente más de un valor con máxima frecuencia; cuando la distribución tiene un único valor más frecuente se denomina unimodal, si presenta dos valores más frecuentes la distribución es bimodal, trimodal si son tres y así sucesivamente.

El cálculo de la moda se puede aplicar a la distribución de cualquier variable, ya sea cualitativa o cuantitativa. En particular, la moda es el único estadístico que puede calcularse de la distribución de una variable medida a nivel nominal.

Cuando la información de la variable se presenta con los datos agrupados en intervalos, entonces hablamos del intervalo modal y podemos dar como valor aproximado más frecuente el punto medio del intervalo (la marca de clase). Por otro lado, con una variable cuantitativa con un gran número de valores, el número de valores repetidos suele ser muy pequeño y no tiene demasiado sentido hablar de moda, es preferible agrupar los datos en intervalos y señalar el intervalo modal, el que tiene máxima frecuencia por unidad de amplitud.

La moda es un estadístico que se le puede pedir al ordenador que nos los determine, no obstante su cálculo no reporta ninguna complejidad, mirando las tablas de frecuencia se obtiene inmediatamente. En las tablas de distribución de frecuencias anteriores podemos constatar que la moda de la variable de estado civil es el valor 1 (estar casado), de la variable sobre la situación económica es 5 (muy mala), de la variables de estudios es 3 (secundaria de 1ª etapa) y de la variable de edad 44 años. Esta información aparece recogida en la Tabla III.3.8. En el caso de la variable del PIB per cápita de los países del mundo, de los 190 de los que se dispone de información, la moda es un valor que tiene poco sentido pues cada valor de la variable tiene frecuencia 1, la distribución es "multimodal"; en la tabla aparece el dato del primer país, el que tiene menor PIB.

3.1.2. *La mediana*

La (*Me*) es el valor de la variable que deja igual número de casos por encima y por debajo de él. Por tanto, es el valor que ocupa la posición central de la distribución de la variable, a partir de éste valor la mitad de los valores son menores o iguales que él y la otra mitad son mayores o iguales. Es el percentil 50 o el segundo cuartil, el que acumula el 50% de los casos, como veremos al comentar esta medida. Puesto que se trata de determinar un centro, el concepto de mediana tan sólo se puede aplicar a las

variables que tengan como mínimo un orden, por lo que no es aplicable a las variables nominales, tan sólo es interpretable a partir de una medición ordinal.

En el cálculo de la mediana tan sólo intervienen las posiciones relativas de los valores de las variables, es decir, el orden de los valores. Por lo que la mediana no se ve afectada por los valores extremos de la variable, aspecto relevante que en el caso de la media como veremos sí se va producir esta influencia.

En el cálculo de la mediana nos podemos encontrar diversas situaciones:

a) Si consideramos directamente los valores observados, para determinar el valor de la mediana debemos disponer en primer lugar de los datos ordenados del valor inferior al superior. Una vez ordenados nos podemos encontrar en dos situaciones, que el número de valores de la variable sea par o impar:

   - Si tenemos un número impar de valores, la mediana es el valor que se encuentra justo en medio. Si tenemos por ejemplo este conjunto de datos sobre el número de veces que se va al cine anualmente de un grupo de 9 individuos:

     1 3 3 4 6 7 7 9 12

     La mediana es el valor 6. Si los datos son estos:

     2 2 2 5 5 5 8 8 8

     La mediana es el valor 5.

   - Si el número de valores es par, se trata de tomar los dos valores que ocupan las posiciones centrales, sumarlos y dividir por 2. Por ejemplo, de estos datos de 10 individuos:

     1 3 3 4 5 7 7 9 10 12

     La mediana es el valor 6 que resulta de calcular: $\frac{5+7}{2}=6$.

     Si los datos son estos:

     2 2 2 5 5 5 5 8 8 8

     La mediana también sería el valor 5 que resulta de: $\frac{5+5}{2}=5$.

b) Si disponemos de los datos en una tabla de frecuencias, podemos calcular la mediana a partir de las frecuencias absolutas o relativas acumuladas.

   - Si el número de casos n es impar, la mediana es el valor de la variable al que le corresponde la primera frecuencia acumulada más grande que $n/2$, o lo que es lo mismo, el primer valor que acumula más del 50%.
     Por ejemplo, con la distribución siguiente:

x_i	n_i	N_i	P_i
1	2	2	22,2
3	3	5	55,6
5	3	8	88,9
9	1	9	100
Total	9		

La segunda frecuencia acumulada que supera $n/2$, es decir, que supera 9/2=4,5, es la frecuencia acumulada 5, que corresponde al valor de mediana 3. Es decir, 3 es el valor de la mediana que tiene el tanto por ciento acumulado superior al 50%.

- Si n es par y ningún valor tiene frecuencia acumulada igual a $n/2$, entonces la mediana es el valor al cual le corresponde la primera frecuencia acumulada más grande que $n/2$, o el primer valor que acumula más del 50%.
  En el ejemplo siguiente:

x_i	n_i	N_i	P_i
1	2	2	20
3	4	6	60
5	3	9	90
9	1	10	100
Total	10		

El valor de la mediana, el que supera a $n/2=5$, y el que también supera el 50%, es la frecuencia acumulada 6 que corresponde al valor 3.

- Si n es par y un valor tiene frecuencia acumulada igual a $n/2$, la mediana es el resultado de sumar este valor y el siguiente y dividir por 2.
  En este ejemplo:

x_i	n_i	N_i	P_i
1	2	2	20
3	3	5	50
5	3	8	80
9	2	10	100
Total	10		

La frecuencia acumulada igual a $n/2=5$ y que acumula el 50% corresponde al valor 4. Para determinar la mediana procedemos a calcular $\frac{3+5}{2}=4$.

c) Si disponemos de los datos agrupados en intervalos, solamente podemos hacer un cálculo aproximado de la mediana, puesto que la agrupación implica la pérdida de la información necesaria para su cálculo preciso. Un cálculo aproximado consistiría simplemente en tomar el intervalo que contiene la mediana, al que le corresponde una frecuencia acumulada mayor que $n/2$, o del 50%, y dar como valor aproximado de la mediana el punto medio o marca de clase del intervalo. Alternativamente se puede estimar un valor intermedio proporcional a la distribución de casos dentro del intervalo en función de la frecuencia acumulada sobrante desde $n/2$.

En los ejemplos de las variables elegidas de la Tabla III.3.8 la mediana de la variable edad es de 46 años, y de 10.338,97 dólares en el caso del PIB per cápita.

3.1.3. *La media*

La media o media aritmética ($\bar{x}$) es un valor que nos proporciona el promedio aritmético de todos los valores de una variable con una escala de medición cuantitativa, un valor central de la distribución de valores de la variable. Es una medida que tiene en cuenta todos los casos y se calcula sumando todos los valores observados (válidos) y dividiendo el resultado por el total de casos, que se expresa mediante la fórmula:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$ *Ecuación 1*

En esta expresión consideramos los valores x_i de todos los casos, uno a uno, y los sumamos entre sí, por tanto, los valores x_i se repiten tal y como están dispuestos en la matriz de datos considerando los n casos o individuos y efectuando en consecuencia n sumas. Si consideramos este ejemplo sencillo, donde disponemos de los valores de las notas de un examen (x_i) de 10 personas: 3 4 5 5 6 6 7 7 8 9 de que disponemos en la tabla siguiente como matriz de datos, con $i=1...n$ y $n=10$[11]:

i	x_i
1	3
2	4
3	5
4	5
5	6
6	6
7	7
8	7
9	8
10	9
Total	60

La media se calcula así:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{3+4+5+5+6+6+7+7+8+9}{10} = \frac{60}{10} = 6$$

Otra forma de obtener el mismo resultado es considerar la tabla distribución frecuencias absolutas n_i de cada valor, es decir, el número de repeticiones de cada valor x_i, y multiplicarlo por éste. Ahora la fórmula se expresa como:

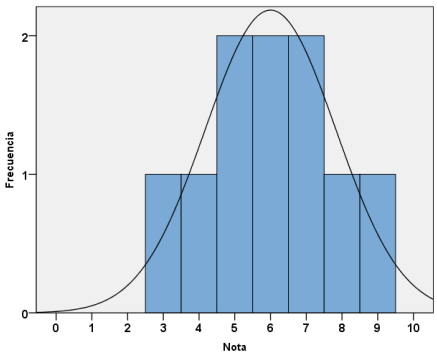
$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \cdots + x_n \cdot n_n}{n} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}$$ *Ecuación 2*

[11] Hemos dispuesto los individuos 1 a 10 de forma ordenada según la nota, 3 a 9, pero cualquier disposición sería igualmente válida.

En este caso los valores x_i de la expresión representan los valores distintos expresados en el subíndice i, donde se consideran los k posibles valores de la variable y se efectúan en consecuencia k sumas, ponderando (multiplicando) cada valor por su frecuencia absoluta.

En el ejemplo anterior la media se calcula a partir de la tabla de frecuencias siguiente que se acompaña del histograma de frecuencias absolutas:

i	x_i	n_i	$x_i \cdot n_i$
1	3	1	3
2	4	1	4
3	5	2	10
4	6	2	12
5	7	2	14
6	8	1	8
7	9	1	9
	Total	10	60



El cálculo es así:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n} = \frac{\sum_{i=1}^{7} x_i \cdot n_i}{10} = \frac{3\cdot 1+4\cdot 1+5\cdot 2+6\cdot 2+7\cdot 2+8\cdot 1+9\cdot 1}{10} = \frac{60}{10} = 6$$

En el caso de la variable edad de la Tabla III.3.8 la media que se obtiene de la población entrevistada es de 48,32 años, mientras que el PIB per cápita medio de los países del mundo es de 16.486,51 dólares.

La media es una medida de resumen de la distribución, es un valor característico que representa el conjunto de valores de la variable, pero nada nos dice de cómo se distribuyen los valores individualmente. Cuando afirmamos por ejemplo que los ingresos medios per cápita del hogar de una sociedad determinada son de 1.000€, damos una información importante e interesante del comportamiento promedio de todos los ingresos de la población que refleja la situación económica de un hogar medio o típico. Evidentemente sabemos que hay familias con ingresos inferiores hasta niveles de pobreza severa, y familias con ingresos superiores hasta niveles de ingresos millonarios. Pero en conjunto, entre los que tienen más y menos, la media nos indica un punto intermedio de esa distribución de ingresos altos y bajos. Cuando introduzcamos las medidas de dispersión podremos precisar esta idea de valores que se sitúan por encima o por debajo de la media[12]. Pero conviene destacar que el valor de la media se ve afectado por la existencia de valores extremos de la variable.

El cálculo de la media también se puede efectuar con las frecuencias relativas f_i, con proporciones o en porcentajes p_i. En este caso las expresiones serían:

[12] Un recurrido ejemplo para expresar esta idea: si tenemos dos personas y un pollo que comer, y resulta que todo el pollo se lo come una de las dos, la media resultante es 1+0/2=0,5, es decir, que se han comido medio pollo cada uno, y nunca más lejos de la realidad. Una media sin la información de su dispersión es insuficiente: no es lo mismo esta situación desigual que un reparto equitativo, si bien las medias que se obtienen son iguales.

$$\bar{x} = x_1 \cdot f_1 + x_2 \cdot f_2 + \cdots + x_n \cdot f_n = \sum_{i=1}^{k} x_i \cdot f_i$$ *Ecuación 3*

$$\bar{x} = \frac{x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n}{100} = \frac{\sum_{i=1}^{k} x_i \cdot p_i}{100}$$ *Ecuación 4*

Con los datos del ejemplo anterior, las distribuciones relativas y los cálculos son los siguientes:

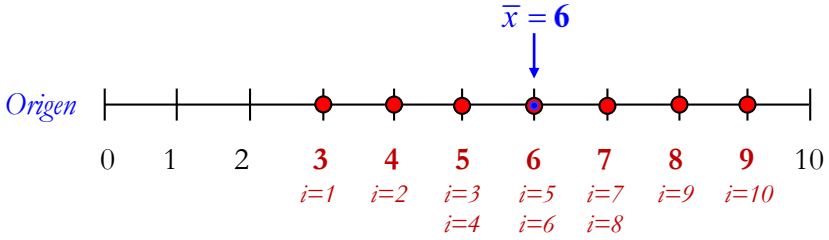
i	x_i	n_i	f_i	$x_i \cdot f_i$	p_i	$x_i \cdot p_i$
1	3	1	0,1	0,3	10	30
2	4	1	0,1	0,4	10	40
3	5	2	0,2	1,0	20	100
4	6	2	0,2	1,2	20	120
5	7	2	0,2	1,4	20	140
6	8	1	0,1	0,8	10	80
7	9	1	0,1	0,9	10	90
Total		10	1	60	100	600

$$\bar{x} = \sum_{i=1}^{7} x_i \cdot f_i = 3 \cdot 0{,}1 + 4 \cdot 0{,}1 + 5 \cdot 0{,}2 + 6 \cdot 0{,}2 + 7 \cdot 0{,}2 + 8 \cdot 0{,}1 + 9 \cdot 0{,}1 = 6$$

$$\bar{x} = \frac{\sum_{i=1}^{7} x_i \cdot p_i}{100} = \frac{3 \cdot 10 + 4 \cdot 10 + 5 \cdot 20 + 6 \cdot 20 + 7 \cdot 20 + 8 \cdot 10 + 9 \cdot 10}{100} = \frac{600}{100} = 6$$

La media es un estadístico que solo se puede utilizar con variables cuantitativas y con todos los datos sin agrupar en intervalos. Cuando los datos están agrupados en intervalos lo único que podemos obtener es una aproximación al valor de la media, puesto que perdemos parte de la información como resultado de la agrupación. En este caso la solución consiste en tomar el valor medio del intervalo y aplicar la fórmula.

El concepto de media se relaciona con la idea de distancia y de punto medio en el espacio. En física ese promedio se corresponde con el centro de masas o el centro de gravedad[13]. La representación siguiente expresa esta idea:



[13] Estas expresiones las utilizaremos en algún momento posterior en este manual.

Además de la media aritmética existen otras formas de calcular la media: la media ponderada, la media geométrica o la media armónica.

La media ponderada ($\bar{x}^p$) es similar a la media aritmética pero en su cálculo se consideran unos pesos w_i que afectan a cada valor de la variable con el objetivo de tener en cuenta la importancia diferente de cada valor. Su expresión general es:

$$\bar{x}^p = \frac{x_1 \cdot w_1 + x_2 \cdot w_2 + \cdots + x_n \cdot w_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i} \qquad \textit{Ecuación 5}$$

Si los pesos w_i fueran constantes e igual a 1 obtendríamos la fórmula de la media aritmética.

Un ejemplo de aplicación sencillo consistiría en obtener la media ponderada de las notas de una asignatura a partir de las puntuaciones o valores de evaluaciones parciales que tienen una valoración distinta. Consideremos que la evaluación se compone de dos partes, unos ejercicios prácticos y un examen final, y se pondera atribuyendo un peso del 30% a los ejercicios y un peso del 70% al examen, o lo que es lo mismo, si se expresan en proporciones, unos pesos de 0,3 y de 0,7. La nota final, la nota media ponderada, implicará multiplicar el resultado de cada nota por su peso, sumarlas y dividir el resultado por la suma de los pesos. Veamos el resultado para una persona que tiene una nota de 8 en los ejercicios y un 7 en el examen.

Si aplicamos el peso en términos de porcentajes, el cálculo es de la media ponderada es:

$$\bar{x}^p = \frac{8 \cdot 30 + 7 \cdot 70}{30 + 70} = \frac{240 + 490}{100} = \frac{730}{100} = 7,3$$

Si aplicamos el peso en términos de proporciones, la media ponderada da lugar al mismo resultado:

$$\bar{x}^p = \frac{8 \cdot 0,3 + 7 \cdot 0,7}{0,3 + 0,7} = \frac{2,4 + 4,9}{1} = 7,3$$

La media ponderada se aplica también al IPC, el Índice de Precios de Consumo. En el caso de INE para España[14] el IPC es una medida estadística de la evolución del conjunto de precios de los bienes y servicios que consume la población residente en viviendas familiares en España. El cálculo del IPC final es la media de los precios de diversos grupos de productos, cada grupo de productos se pondera con un peso diferente en función del gasto realizado a partir de la información que se obtiene de la Encuesta de Presupuestos Familiares (EPF) y de otras fuentes. Las ponderaciones de cada artículo representan la relación entre el gasto realizado en las parcelas representadas por dicho artículo y el gasto total realizado en todas las parcelas cubiertas por el índice.

[14] Se puede consultar en la página web del Instituto Nacional de Estadística (INE) de España: http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176802&menu=ultiDatos&idp=1254735976607, en particular la metodología base 2011: http://www.ine.es/prensa/np701.pdf.

En la siguiente tabla se incluye el peso de cada uno de los 12 grandes grupos y su comparación con los pesos vigentes hasta el año 2011:

Ponderaciones de grupos (tanto por cien)

Grupo	2011	2012	%
01. Alimentos y bebidas no alcohólicas	18,16	18,26	0,6
02. Bebidas alcohólicas y tabaco	2,87	2,89	0,7
03. Vestido y calzado	8,59	8,34	-2,9
04. Vivienda	11,70	12,00	2,6
05. Menaje	6,84	6,67	-2,5
06. Medicina	3,21	3,14	-2,1
07. Transporte	14,74	15,16	2,9
08. Comunicaciones	3,98	3,85	-3,3
09. Ocio y Cultura	7,64	7,54	-1,3
10. Enseñanza	1,38	1,42	2,8
11. Hoteles, cafés y restaurantes	11,52	11,46	-0,5
12. Otros bienes y servicios	9,37	9,26	-1,2
TOTAL	**100**	**100**	

Fuente: INE

El uso de la media ponderada también se puede aplicar en el caso de querer calcular una media a partir de otras medias. Si en una subpoblación con n_1 individuos se obtiene una media $\overline{x}_1$, y en otra subpoblación con n_2 individuos se obtiene una media $\overline{x}_2$, entonces la media conjunta de las dos subpoblaciones es una media ponderada que se calcula mediante:

$$\overline{x}^p = \frac{\overline{x}_1 \cdot n_1 + \overline{x}_2 \cdot n_2}{n_1 + n_2} \qquad \textit{Ecuación 6}$$

En el caso que consideremos cualquier número de medias k, la generalización de la expresión anterior es:

$$\overline{x}^p = \frac{\overline{x}_1 \cdot n_1 + \overline{x}_2 \cdot n_2 + \cdots + \overline{x}_k \cdot n_k}{n_1 + n_2 + \cdots + n_k} \qquad \textit{Ecuación 7}$$

Veamos un caso concreto. Según la encuesta de población activa del tercer trimestre del año 2015, y para datos a nivel estatal, sabemos que el número medio de horas efectivas trabajadas por los ocupados varones es de 40,4 horas semanales, mientras que para las mujeres ocupadas la media es de 34,5 horas semanales. Para conocer cuál el número medio de horas trabajadas conjuntamente por varones y mujeres a partir de esta información necesitamos ponderar las medias por el número de varones ocupados (9.896.500) y de mujeres ocupadas (8.152.200) obteniendo:

$$\overline{x}^p = \frac{40,5 \cdot 9896500 + 35,4 \cdot 8152200}{9896500 + 8152200} = \frac{681069500}{18048700} = 37,7$$

Un error del inexperto/a, relativamente común, consiste en tomar directamente las dos medias, sumarlas y dividir por 2. Si lo hiciéramos así obtendríamos una media de 37,4, un valor erróneo que difiere en tres décimas. Se podría decir que la diferencia no es muy importante, piénsese si tenemos en cuenta lo que implicaría en dinero una hora semanal multiplicada por las casi 18 millones de personas ocupadas. En ese cálculo erróneo que no se ha considerado la ponderación, de hecho, hemos supuesto que el número de varones y de mujeres ocupados/as era el mismo, circunstancia que no obedece a la realidad del mercado de trabajo en España. Por tanto el cálculo será más

erróneo cuantas más diferencias de peso se den entre los grupos considerados o el número de medias que se comparan.

En el apartado de análisis exploratorio que trataremos más delante veremos que existe un cálculo de la media que recibe el nombre de media recortada (o truncada, $\bar{x}^r$). Se obtiene de la misma forma que la media aritmética pero después de eliminar un determinado porcentaje de casos atípicos por abajo y por arriba de la distribución. Precisamente este cálculo compensa la posible influencia de comportamientos extremados de algunos individuos que distorsiona el cálculo y ante los que la media es sensible, y una vez eliminados el indicador refleja mejor el comportamiento central mayoritario y se dice que el estimador es más robusto. Se puede eliminar entre un 5% o un 25% de los casos extremos. Veamos un sencillo ejemplo. Si disponemos de las siguientes observaciones para estimar el número de veces que los clientes frecuentan un local de ocio durante el año: 7 8 9 9 10 10 11 11 12 63, se obtiene una media aritmética de 15 veces, si recortamos la distribución al 10% eliminaremos el primero y el último caso, resultando ahora que la media es 10, un valor más real que el anterior dado el conjunto mayoritario de observaciones del ejemplo.

La media aritmética, la media ponderada y la media recortada son las más utilizadas. Existen otros cálculos de media como la media geométrica ($\bar{x}^g$) que se calcula multiplicando todos los valores y hallando la raíz con exponente n (el número de casos):

$$\bar{x}^g = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \qquad \textit{Ecuación 8}$$

Tiene la ventaja también de se menos sensible a la presencia de valores extremos, si bien su interpretación no es tan intuitiva. Se emplea cuando varias cantidades son multiplicadas para producir un total. Por ejemplo, si la tasa de desempleo aumentó en el momento t_1 un 2% y un 8% en t_2, el promedio anual de crecimiento del desempleo no es la media aritmética $\bar{x} = \frac{2+8}{2} = 5$, sino la media geométrica $\bar{x}^g = \sqrt[2]{2 \cdot 8} = 4$.

La media geométrica solamente se puede aplicar a variables de razón y es necesario que todos los valores sean mayores que cero. La media geométrica es siempre menor o igual que la media aritmética.

La media armónica ($\bar{x}^h$) es la media que se obtiene de dividir el número de casos entre la suma de la inversa de los valores que se promedian:

$$\bar{x}^h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} \qquad \textit{Ecuación 9}$$

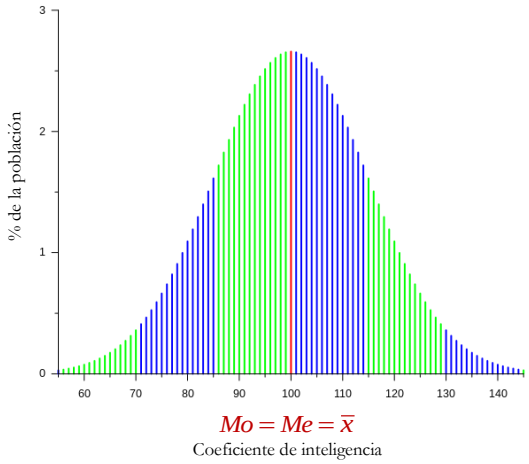
Se emplea cuando se manejan velocidades o ratios. La media armónica resulta poco influida por valores mucho más grandes que el resto pero en cambio es sensible a valores mucho más pequeños que el conjunto. Se utiliza solamente con variables de razón y es necesario que todos los valores sean distintos de cero. La media armónica siempre es menor o igual que la media aritmética.

3.1.4. *Comparación entre la media, la mediana y la moda*

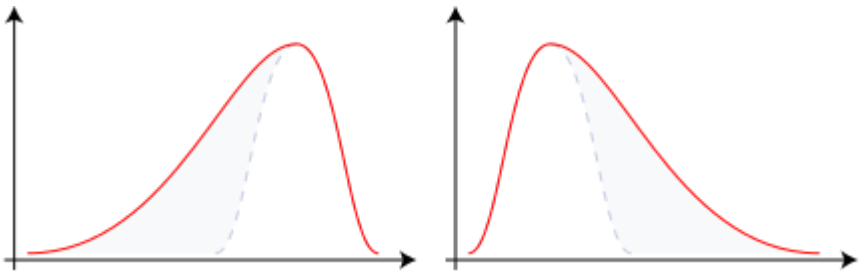
Cuando analizamos los datos estadísticos para determinar el valor central de la distribución es de interés utilizar simultáneamente las distintas medidas, la moda, la mediana y la media. Estos valores puedes ser próximos entre sí, en ese caso estamos ante una distribución que se caracteriza por tener un comportamiento simétrico entre los casos que están por encima o por debajo de los valores centrales. Si difieren entre sí se debe a la mayor concentración en los valores bajos o altos de la variable.

El cociente o coeficiente intelectual (CI) es una medida valora con puntuaciones el nivel de inteligencia de las personas como resultado de la aplicación de unos test estandarizados que generan distribuciones simétricas donde las medidas de tendencia central coinciden en el valor 100.

Gráfico III.3.10. Histograma del coeficiente de inteligencia



Las distribuciones perfectamente simétricas contienen una moda y tienen el mismo valor para la media, la mediana y la moda. Posteriormente determinaremos una medida de resumen de la simetría, o de la asimetría que suele ser la situación más habitual, es decir, distribuciones donde las tres medidas de tendencia central se ubican en posiciones distintas. En esos casos la distribución se dice que está sesgada, hacia la derecha o hacia la izquierda. En una distribución sesgada hacia la derecha (positivamente sesgada) la media es el valor más alto, por estar afectada por la presencia de valores más extremos, quedando la mediana a la izquierda y más allá la moda.



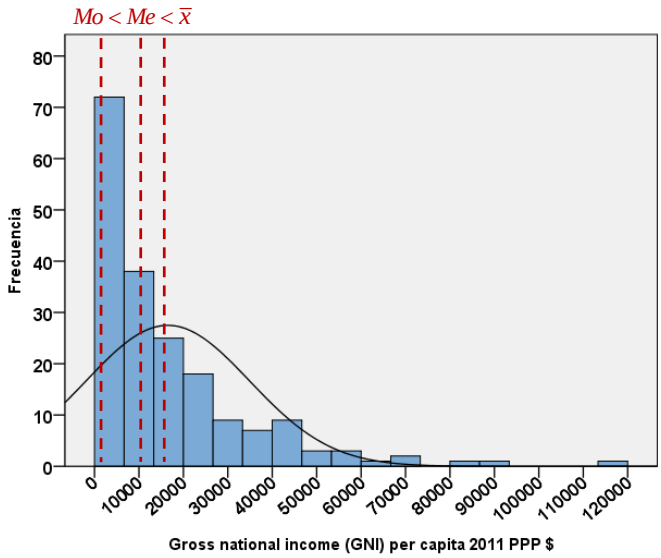
$\bar{x} < Me < Mo$
Sesgo a la izquierda

$Mo < Me < \bar{x}$
Sesgo a la derecha

De forma equivalente, en una distribución sesgada hacia la izquierda (negativamente sesgada) la media es el valor más bajo y a su derecha quedan la mediana y la moda.

En el Gráfico III.3.10 se reproduce, con los datos de la matriz IDH2014, el histograma con la distribución de las variables del PIB per cápita de los países del mundo. Se observa claramente cómo se da una concentración de países en los niveles bajos de riqueza, siendo la riqueza de algunos países un comportamiento extremo por la derecha de la distribución. Se produce así un sesgo positivo donde la media del PIB per cápita es de 16.487 dólares, por encima de la mediana de 10.339 dólares y de una moda 444 dólares.

Gráfico III.3.11. Histograma del producto interior bruto per cápita de los países del mundo



Fuente: Naciones Unidas, 2014

Cuando la población está sesgada negativa o positivamente, con frecuencia la mediana resulta ser la mejor medida de posición central, debido a que siempre está entre la moda y la media. La mediana no se ve altamente influida por la frecuencia de aparición de un solo valor como es el caso de la moda, ni se distorsiona con la presencia de valores extremos como la media.

3.2. Medidas de posición no central

3.2.1. *Valores extremos*

Los valores extremos son el mínimo y el máximo de la distribución de los datos de una variable cuantitativa, nos indican desde qué valor se inicia la distribución y hasta qué valor llega, posiciones no centrales relevantes de una variable.

En el caso de la variable de edad los datos observados van desde los 18 años hasta los 94 (ver). En el caso de la variable del PIB per cápita oscila entre el país más pobre con

443,96 dólares (la República Democrática del Congo) al país más rico con 119.029,12 dólares (Catar).

3.2.2. *Percentiles*

Los percentiles (P_k) son otra forma de resumir los datos para dar cuenta de la posición de la distribución cuando disponemos de variables medidas como mínimo a nivel ordinal, aunque su mayor interés se da cuando las variables son cuantitativas. El percentil es una medida de posición que nos proporciona el valor de la variable x_i que acumula un determinado porcentaje de casos k. Por lo tanto, la mediana no es más que un caso particular de percentil, el percentil 50 (P_{50}), aquel valor que acumula el 50% de los casos y divide a la distribución en dos partes iguales. Pero podemos extender este cálculo a cualquier valor porcentual, existen 99 valores que dividen una distribución en 100 partes iguales. Por ejemplo, el percentil 20 de la variable de ingresos corresponde al valor que marca el 20% más pobre de la población, el 20% de los que tienen unos ingresos por debajo o igual a ese valor.

Si en lugar de emplear porcentaje empleáramos proporciones hablaríamos de centiles (CN_k), es decir, el valor de la variable x_i tal que la proporción k de casos es menor o igual a x_i. Si k es 50, el percentil 50 (P_{50}) se corresponde con el centil 0,5 ($CN_{½}$).

Es habitual calcular los percentiles que se corresponden con los cuartos de la distribución. En este caso se denominan cuartiles. Los cuartiles (C_k) son tres: el primer cuartil (C_1) es el valor de la variable tal que el 25% de los casos son menores o iguales que él, y el 75% restante son mayores, y coincide con el percentil 25. El segundo cuartil (C_2) coincide con la mediana y el tercer cuartil (C_3), o percentil 75, es el valor de la variable tal que el 75% de los casos son menores o iguales que él, y el 25% restante son mayores.

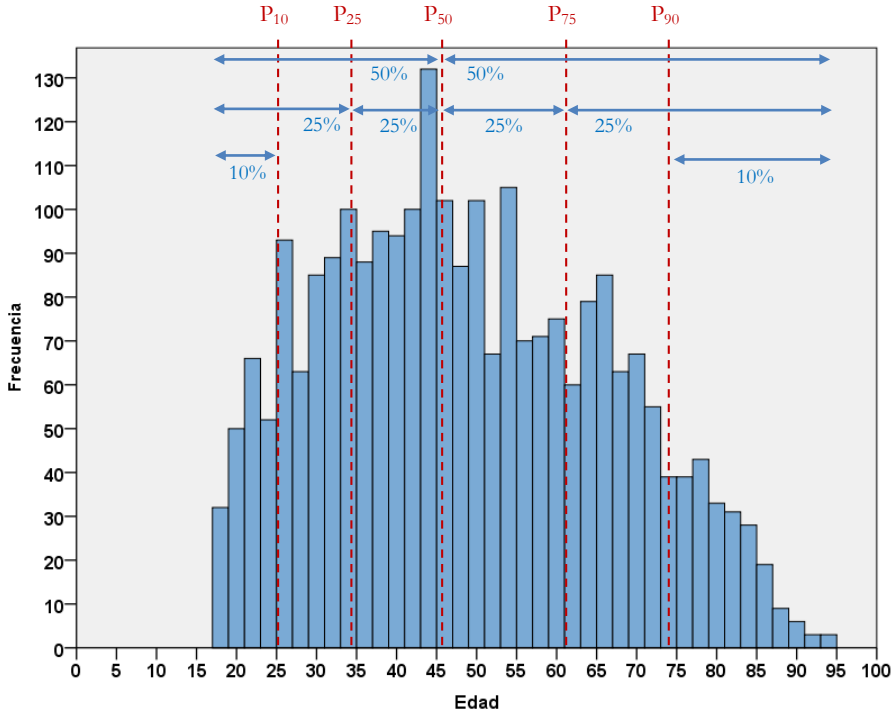
De forma análoga se definen los quintiles (Q_k) y los deciles (D_k) a partir de la división de la distribución en cinco y nueve partes, respectivamente. Desde el primer quintil (Q_1) que corresponde al 20%, tenemos sucesivamente el 40% (Q_2), el 60% (Q_3) y el 80% (Q_4). El primer decil (D_1) se corresponde con el percentil 10, y de 10 en 10, se completan hasta el decil 9 (D_9) que corresponde al percentil 90. A los cuartiles, quintiles, deciles y cuantiles se denominan en general cuantiles.

En la Tabla III.3.9 se presentan estas diferentes medidas calculadas para la variable edad del Barómetro del CIS y en el histograma del Gráfico III.3.12 se puede visualizar su posición.

Tabla III.3.9. Percentiles de la variable P32 (Edad)

Percentiles										
10	20	25	30	40	50	60	70	75	80	90
25	32	34	37	42	46	52	58	62	65	73

Gráfico III.3.12. Histograma de la variable P32 (Edad) con los percentiles



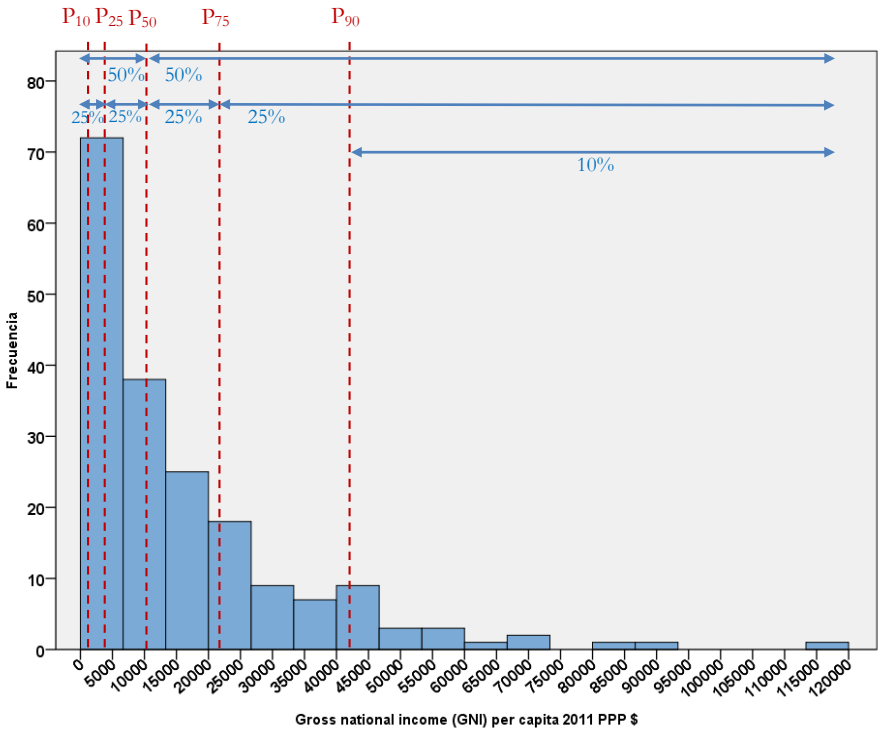
Se puede apreciar por ejemplo cómo el 50% de las personas más jóvenes, las que tienen entre 18 y 46 años, tienen un recorrido de 27 años entre esas dos edades mientras que el 50% de más edad de la población, entre 46 y 95 años, el recorrido es mayor, de 49 años. Es decir, que para alcanzar el mismo número de personas (la mitad) en un caso se necesitan menos valores de edad porque los individuos están concentrados en las edades más jóvenes, mientras las edades mayores se dispersan más, se necesita recorrer más edades para alcanzar el mismo número de casos, la otra mitad de las personas entrevistadas. En otras palabras, el 50% de las dos áreas que marca la mediana, en un caso, entre los más jóvenes, tiene una base más pequeña y más altura mayor, y en el otro, el de los mayores, tiene una base mayor y una altura menor. Parecidos comentarios podrían realizarse teniendo en cuenta los cuartiles o los deciles poniendo de manifiesto los lugares de la distribución (los percentiles) donde se da mayor concentración o dispersión de valores.

En el 0 vemos de nuevo el histograma del producto interior bruto per cápita de los países del mundo donde se puede apreciar la elevada concentración de países en los niveles bajos de este indicador de riqueza. Se han destacado como anteriormente algunos percentiles de la Tabla III.3.10.

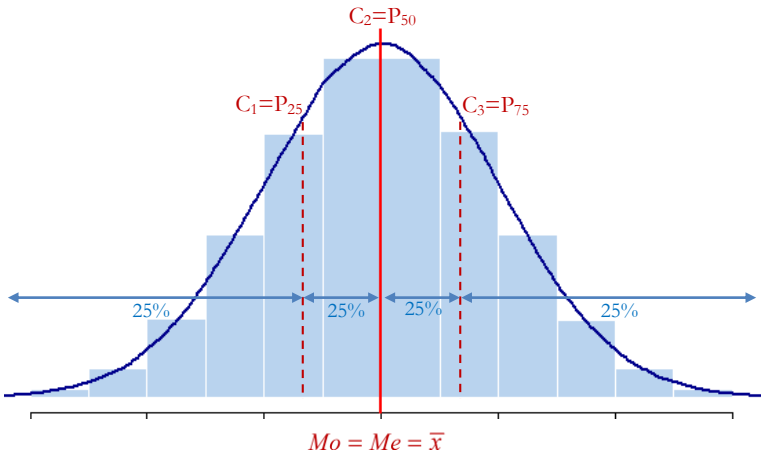
Tabla III.3.10. Percentiles del PIB per cápita de los países del mundo

Percentiles										
10	20	25	30	40	50	60	70	75	80	90
1499,93	2778,99	3629,59	4897,50	7028,65	10338,97	14167,79	19315,93	22025,64	25333,87	41850,53

Gráfico III.3.13. Histograma del producto interior bruto per cápita de los países del mundo con los percentiles



Si la distribución de la variable fuese perfectamente simétrica, además de coincidir moda, mediana y media, tendríamos una distribución de percentiles con los mismos recorridos de valores a uno y otro lado de los valores centrales. En el gráfico siguiente se marcan los valores correspondientes a los cuartiles.



Así pues, el uso de los percentiles nos permite posicionar con pocos valores toda la distribución. Veremos que algunas posiciones se consideraran de especial relevancia en el análisis estadístico. Por ejemplo, entre el primer y tercer cuartil se encuentra el 50% de los datos. Entre el percentil 2,5 y el percentil 97,5 se encuentra el 95% de los

datos. Esta idea la tendremos en consideración a menudo en el análisis de datos estadísticos.

Como sucedía en el caso de la mediana, cuando disponemos de los datos agrupados en intervalos tan sólo podemos dar un valor aproximado del percentil a partir de la marca de clase del intervalo.

3.3. Medidas de dispersión

Las medidas de dispersión o de variabilidad constituyen una información fundamental para complementar la de centralidad y dar cuenta del comportamiento de una distribución. Como sugerimos anteriormente, las medidas de tendencia central nos proporcionan un tipo de información en forma resumida que resulta insuficiente para caracterizar la naturaleza de una distribución.

Un ejemplo sencillo ilustrará lo que decimos. Si consideramos dos personas con un salario mensual de 500 y 4500 euros, y calculamos la media, obtenemos un valor de 2500 euros, es decir, en promedio, el salario está bastante por debajo de la persona que más gana cada mes y bastante por encima de la persona que menos gana. El valor de la media esconde la existencia de importantes diferencias de ingresos entre estas dos personas. Por otro lado, supongamos que las dos personas cobran un salario de 2400 y de 2600 euros al mes, en este caso la media vuelve a ser el mismo valor que en el caso anterior, 2500 euros, pero en este caso las diferencias entre estas dos personas no son apenas importantes. Por tanto, el grado de "representatividad" del valor de la media es muy distinto y si no informamos de la diferente dispersión de los datos que la generan no reflejaremos con precisión las características de los datos.

Para solucionar este tipo de problemas podemos recurrir a algún tipo de medida que nos informe sobre las diferencias que producen entre los distintos valores particulares y el valor que proporciona la media. Se trata de evaluar a través de una medida la mayor o menor dispersión del conjunto de valores de una distribución con respecto al valor medio. Un valor alto de esta medida indicará que existen importantes diferencias entre, por ejemplo, personas que cobran altos salarios y bajos salarios que se alejan del salario medio; por el contrario, un valor pequeño de esta medida será indicativo de una distribución de salarios muy parecidos entre sí y cercanos al valor de la media.

Pero existen diferentes tipos de cálculos que nos permiten obtener información sobre la concentración o dispersión de una distribución, entre los más habituales se encuentran el rango, el rango intercuartil y la varianza (junto la desviación típica).

3.3.1. *Rango*

El rango (recorrido o amplitud) es la diferencia entre el valor más grande y el más pequeño de la distribución de una variable:

$$R = x_{\max} - x_{\min}$$ *Ecuación 10*

El rango de la variable edad que aparece en la Tabla III.3.8 es de 76, la diferencia entre 18 y 94. En el caso del PIB per cápita la diferencia entre el país más rico con 119.029,12 dólares y el más pobre con 443,96 dólares es de 118.585,16 dólares.

3.3.2. *Rango intercuartil*

El rango o desviación intercuartil (o intercuatílico) es la diferencia entre el valor del tercer cuartil (percentil 75) y del primer cuartil (percentil 25):

$$RI = C_3 - C_1 \qquad \textit{Ecuación 11}$$

Por tanto es una medida que corresponde al recorrido entre los valores que corresponden al 50% central de la distribución. Es un estadístico que se emplea en el análisis exploratorio de datos como veremos posteriormente.

En los ejemplos de la edad y del PIB per cápita que comentamos al hablar de los percentiles pudimos ver esta información que ahora se convierte en una medida al calcular la diferencia entre el tercer y primer cuartil: en el caso de la edad es la diferencia entre 62 y 34, un rango de 28 años; en el caso del PIB el rango intercuartil es de 18.396,05 dólares, el resultado de la resta entre 22.025,64 y 10.338,97.

3.3.3. *La varianza y la desviación típica*

La principal medida para evaluar el grado de concentración o de dispersión de les datos es la varianza. Se trata de un concepto y un cálculo fundamental en el análisis estadístico pues la variabilidad existente en los datos será lo que centrará el interés del investigador/a: intentar explicar por qué se dan esas diferencias o variaciones. Para ello se calcula una medida que evalúa las diferencias entre los valores particulares los individuos en una distribución y el valor de la media.

El cálculo del estadístico se puede formular de diversas formas en función de si tratamos con datos poblacionales o muestrales, o en función de si lo calculamos con la matriz de datos o a partir de tablas de frecuencias.

a) Con datos poblacionales, referidos al número de casos N de la población, definimos la varianza poblacional (σ^2) como la media de los cuadrados de las desviaciones de los valores de la variable (x_i) en relación a la media (μ):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad \textit{Ecuación 12}$$

Es decir, realizamos tantas sumas como individuos tenemos (N) de la diferencia (la desviación o distancia) entre el valor en la variable de cada individuo y el valor de la media. Este cálculo inicial lo elevamos al cuadrado para eliminar el signo, pues en caso contrario la suma que hiciéramos de las diferencias sin elevar daría cero, y el resultado de la suma lo dividimos entre el número de casos, es decir, lo repartimos, calculando así el promedio de las desviaciones al cuadrado.

Por tanto, podemos expresar este cálculo diciendo:

$$\sigma^2 = \frac{\text{Suma de los cuadrados de las diferencias}}{N} = \frac{SCD}{N} \qquad \textit{Ecuación 13}$$

b) Con datos muestrales, y de forma equivalente pero referidos al número de casos n de la muestra, definimos la varianza muestral (s^2) como la media de los cuadrados de las desviaciones de los valores de la variable (x_i) en relación a la media ($\bar{x}$), pero ajustando el denominador con $n-1$[15]:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad \textit{Ecuación 14}$$

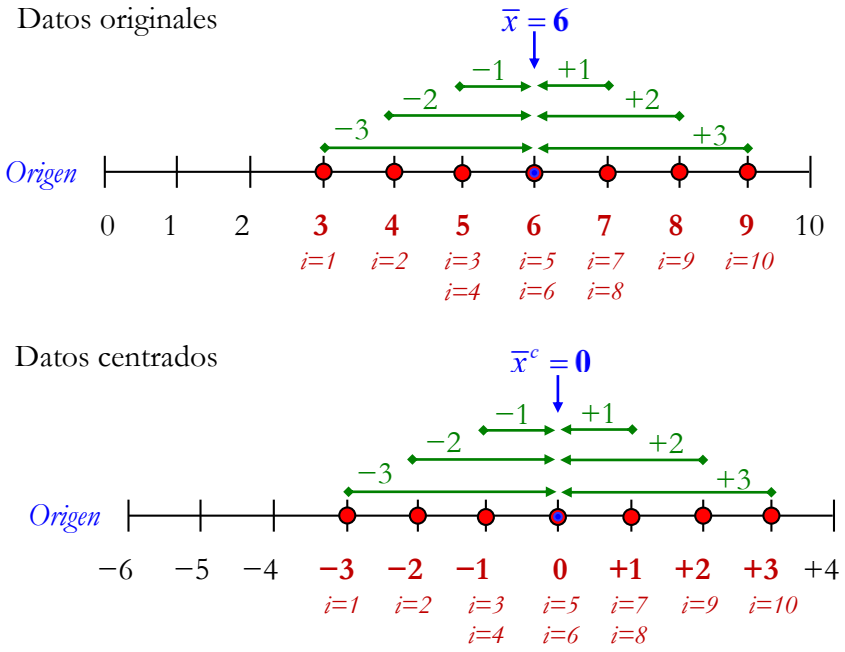
En estas fórmulas, al realizar el cálculo de la diferencia $x_i - \bar{x}$, estamos evaluando la desviación o la dispersión, el alejamiento o la distancia entre un valor concreto y el valor central de la media. Las diferencias pueden ser positivas, indicando que el dato está por encima de la media, o negativas, lo que indica que el valor es inferior a la media. Si la diferencia es cero no existe dispersión, el valor de un individuo coincide con el de la media. Por tanto, si ello sucediera en todos los casos, la varianza sería cero. Así pues la varianza oscila entre un valor mínimo cero y un valor superior que en cada caso será distinto dependiendo de la magnitud de la media y de las unidades, de la unidad de medida y de la magnitud de las desviaciones. La varianza es por tanto una medida de dispersión absoluta. Una distribución con menor desviación típica no es necesariamente una distribución que es menos variable o dispersa que otra. Para ello necesitaremos una medida de dispersión relativa como veremos posteriormente.

Para ilustrar con datos concretos el concepto y el cálculo de la varianza recuperemos el sencillo ejemplo que vimos en el cálculo de la media de las notas de un examen. Siendo la media 6, todas las notas por debajo de ella dan diferencias negativas, cero si son iguales y positivas si están por encima, siendo la suma de todas ellas igual a cero:

		D	CD
i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	3	-3	9
2	4	-2	4
3	5	-1	1
4	5	-1	1
5	6	0	0
6	6	0	0
7	7	1	1
8	7	1	1
9	8	2	4
10	9	3	9
Total	60	$SD=0$	$SCD=30$

[15] La razón que justifica dividir por $n-1$ es estadística, en la medida en que la varianza muestral estima la varianza poblacional es necesario que se dé una propiedad de todo estimador: que sea insesgado (o centrado), es decir, que cuando se calcula se "espera" obtener el valor poblacional, y al dividir por $n-1$ se consigue. De todas formas con muestras grandes las diferencias son insignificantes.

Es decir con este primer cálculo transformamos la nota inicial para expresarla en relación a la nota media: la persona que obtiene un 4 ahora decimos que ha sacado una nota dos puntos por debajo de la media, y la persona que obtuvo un 9 sacó 3 puntos más que la media. Esta operación se denomina centrar los datos y supone que su media sea cero (0 es el valor que corresponde a la media 6 sin centrar). Gráficamente podemos representarlo entendiendo que la desviación de un individuo respecto de la media es una "distancia al centro", de la forma siguiente, con los datos originales y con los datos centrados:



Para el cálculo de la varianza se elevan las diferencias al cuadrado (CD), y evitar así que valga cero (la distancia ahora es cuadrática), y calcula la suma de las diferencias al cuadrado (SCD), el numerador de la varianza, 30. Si los datos fueran poblacionales dividiríamos por 10 y la varianza sería 3, si consideramos que son muestrales dividiremos por $n-1$, es decir, por 9[16]:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{10}(x_i - \bar{x})^2}{10-1} = \frac{SCD}{9} = \frac{30}{9} = 3{,}3$$

El resultado 3,3 se expresa en unidades de nota al cuadrado. Para restituir la unidad de nota sin elevar al cuadrado es preciso calcular la raíz cuadrada obteniendo así la desviación típica o estándar (s):

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad \textit{Ecuación 15}$$

[16] El software estadístico siempre calcula la varianza muestral dividiendo por $n-1$.

En el ejemplo un valor de: $s=\sqrt{3,3}=1,8$, valor que se interpreta como la dispersión promedio de todos los valores: la media de las desviaciones es 1,8 entre las personas que sacaron una nota superior o inferior a la media.

Podríamos interpretarlo también de otra forma. Si tuviéramos que estimar la nota que ha obtenido cada individuo y no tuviéramos más información que la nota media del grupo, podríamos realizar una estimación asignando a cada persona el valor de la media. Evidentemente si no obtienen todos la misma nota nos equivocaremos en nuestra predicción, precisamente el error promedio que cometeríamos al asignar a todos los casos el valor de la media sería la desviación típica.

Estos mismos cálculos se pueden reproducir a partir de la tabla de frecuencias. En este casos las fórmulas de la varianza poblacional y muestral se expresan así:

$$\sigma^2=\frac{\sum_{i=1}^{k}(x_i-\mu)^2\cdot N_i}{N} \qquad \textit{Ecuación 16}$$

$$s^2=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2\cdot n_i}{n-1} \qquad \textit{Ecuación 17}$$

Y los cálculos de esta forma:

			D		*CD*
i	x_i	n_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2\cdot n_i$
1	3	1	-3	9	9
2	4	1	-2	4	4
3	5	2	-1	1	2
4	6	2	0	0	0
5	7	2	1	1	2
6	8	1	2	4	4
7	9	1	3	9	9
Total	60		*SD*=0		*SCD*=30

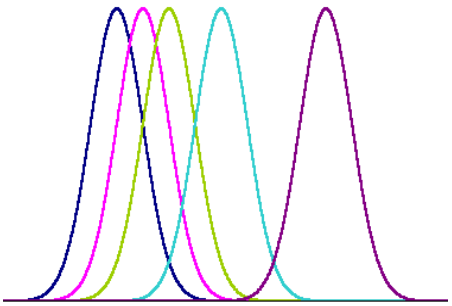
Igualmente se obtiene una varianza de 3,3 y valor medio de las desviaciones de 1,2:

$$s^2=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2\cdot n_i}{n-1}=\frac{\sum_{i=1}^{10}(x_i-\bar{x})^2\cdot n_i}{10-1}=\frac{SCD}{9}=\frac{30}{9}=3,3 \quad\Rightarrow\quad s=\sqrt{s^2}=\sqrt{3,3}=1,8$$

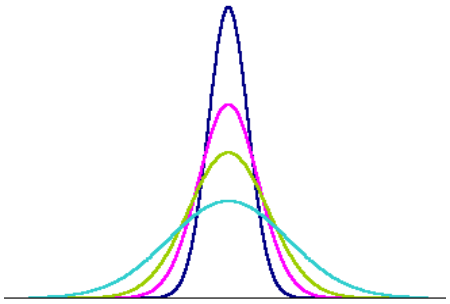
Si consideramos las medidas de dispersión de las variables edad y PIB per cápita de la Tabla III.3.8 podemos ver que el primer caso la desviación típica es de 17,49 años, es decir, que en promedio, entre las personas mayores que la media (48,32) y las personas más jóvenes que la media, la desviación es de 17,49 años 18.383 dólares.

Finalmente podemos considerar la comparación de la varianza y la media. En el gráfico adjunto se puede ver en el gráfico (a) cómo la forma de la distribución se mantiene igual fijando la desviación típica y haciendo variar la media lo que provoca un

desplazamiento horizontal de toda la distribución. En el gráfico (b) se mantiene la media y se hace variar la desviación típica generando una familia de curvas que se hacen más planas a medida que la dispersión aumenta.



(a) Distribuciones con igual desviación típica y diferentes medias



(b) Distribuciones con diferentes desviaciones típicas e igual media

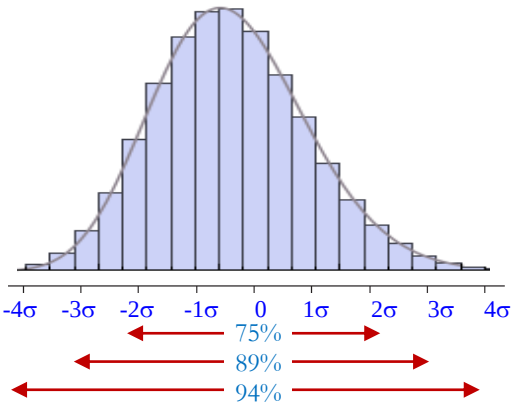
3.3.4. *La desigualdad de Chévichev*

La desviación estándar nos permite determinar, con un buen grado de precisión, dónde están localizados los valores de una distribución de frecuencias con relación a la media. La desigualdad de Chévichev (o de Chebyshov) nos permite constatar cierto comportamiento de regularidad estadística de interés. A partir de cualquier variable numérica, sea cual sea la distribución de los datos, las observaciones se sitúan en torno a la media de la siguiente forma:

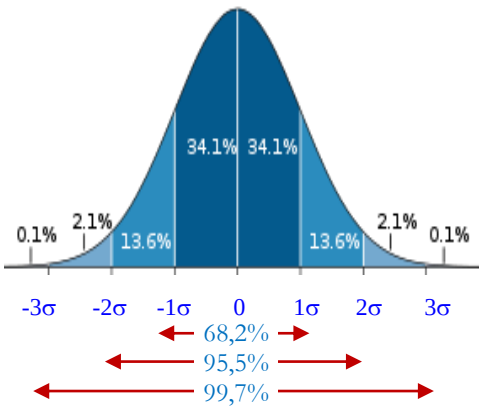
- Como mínimo, el 75% de los casos están en ± 2 desviaciones típicas de la media, es decir, entre $x-2s$ y $x+2s$.
- Como mínimo, el 89% de los casos están en ± 3 desviaciones típicas de la media, es decir, entre $x-3s$ y $x+3s$.
- Como mínimo, el 94% de los casos están en ± 4 desviaciones típicas de la media, es decir, entre $x-4s$ y $x+4s$.

Si la distribución cumple además las condiciones de la distribución normal, como veremos en el próximo capítulo, estos porcentajes se pueden precisar todavía más: el 95,45% está entre ± 2 desviaciones típicas y el 99,87 entre ± 3 desviaciones típicas.

Gráfico III.3.14. Desigualdad de Chévichev

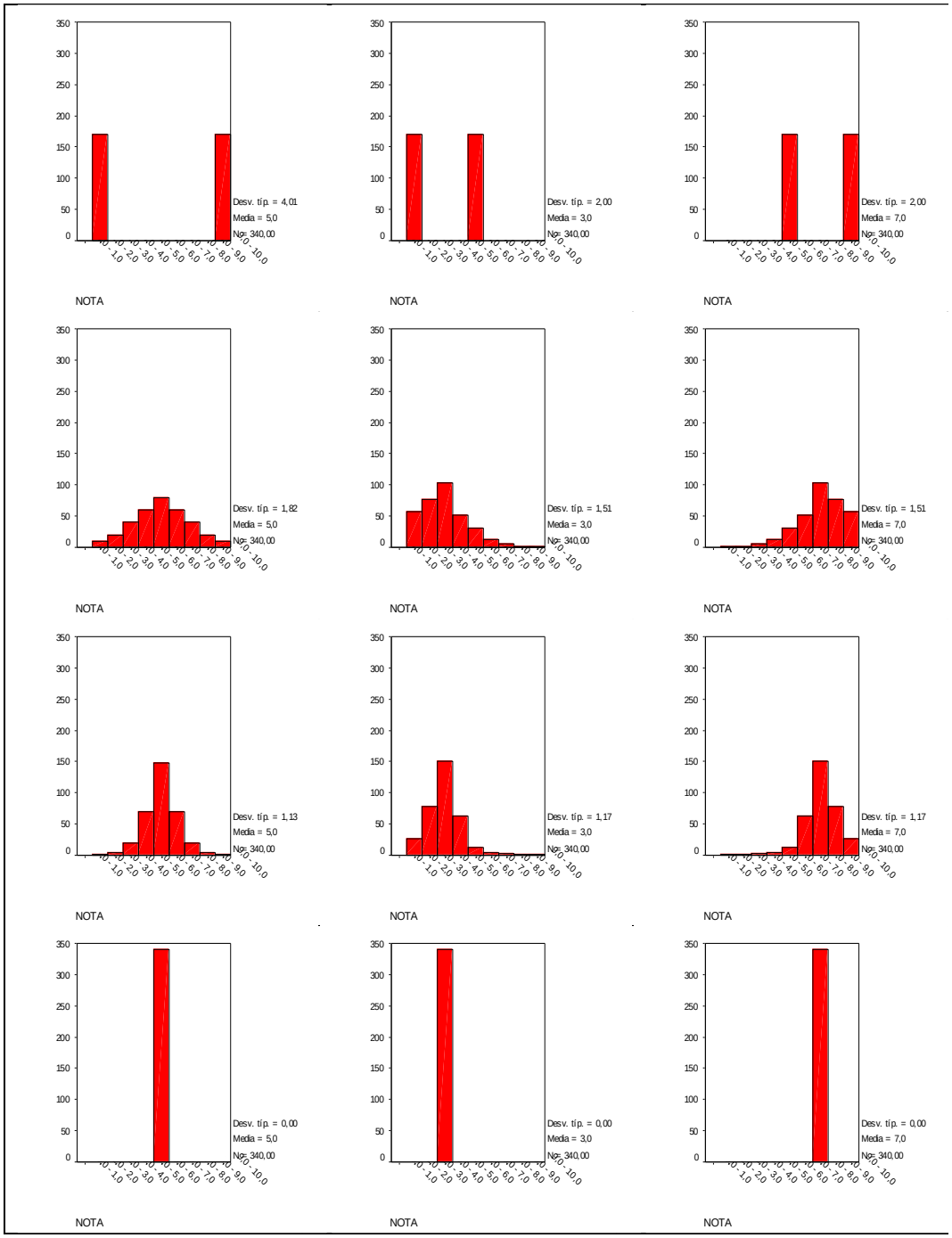


(a) Cualquier distribución



(b) Distribución normal

En los gráficos siguientes y con el ejemplo de las notas de un examen, se distinguen tres columnas donde varía la media entre los 3 (con distribución simétrica), 5 (distribución asimétrica a la derecha) y 7 (distribución asimétrica a la izquierda), y en cada caso cuatro comportamientos de la desviación entre dos situaciones extremas: el grupo se divide en dos valores de nota o todos obtienen la misma nota.



3.3.5. *Dispersión relativa*

La desviación típica es un estadístico de dispersión que se calcula en relación a la media de una variable y se expresa en las unidades de medida de ésta. Si queremos comparar entre sí las desviaciones de distribuciones de distintas variables debemos tener en cuenta que la unidad de medida sea comparable y que la magnitud de las medias tenga características similares, en caso contrario, la comparación carece de sentido. Mediante las medidas de dispersión relativa podemos evitar estos inconvenientes y posibilitar la comparación pues son medidas que estandarizan las unidades. Para ello se divide una medida de dispersión entre una medida de centralidad y al hacerlo las unidades se cancelan.

La medida de dispersión relativa más utilizada es el coeficiente de variación que se define como el cociente entre ls desviación típica y la media, es decir, como proporción de la media aritmética:

$$CV = \frac{s}{\bar{x}} \qquad \textit{Ecuación 18}$$

El coeficiente se suele multiplicar por 100 y se expresa como tanto por ciento, si bien puede superarlo. Solamente se puede emplear con variables positivas y tiene la propiedad de ser invariante ante cambios de escala.

Podemos comparar la dispersión de las variables edad y PIB per cápita calculando sus coeficientes de variación. A partir de los datos de la Tabla III.3.8:

$$CV(edad) = \frac{s}{\bar{x}} = \frac{17,49}{48,32} = 0,362$$

$$CV(PIB) = \frac{s}{\bar{x}} = \frac{18383}{16486,51} = 1,11$$

Se concluye como ya tuvimos ocasión de ver anteriormente que la distribución de la edad es bastante menos dispersa en relación a la distribución de la riqueza.

Otra medida de dispersión relativa es el rango intercuartil relativo que se calcula dividiendo el rano intercuartil entre la mediana:

$$RIR = \frac{C_3 - C_1}{Me} \qquad \textit{Ecuación 19}$$

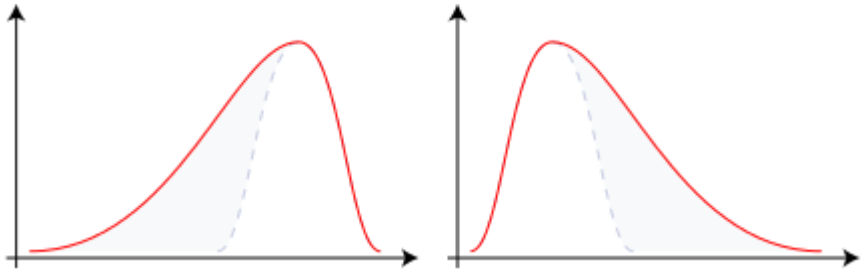
3.4. Medidas de forma

Finalmente la caracterización de una distribución se puede completar con las medidas de forma: la simetría y la curtosis.

3.4.1. *Simetría*

La simetría es una característica que se puede establecer claramente con una representación gráfica como vimos en algún momento anterior, y también se puede

determinar con un cálculo numérico a partir de distribuciones que sean unimodales. La simetría perfecta implica que la moda, la mediana y la media coincidan: $Mo = Me = \bar{x}$. Si se da asimetría los tres indicadores difieren. En el caso de la asimetría positiva se da un sesgo por la derecha, los casos más extremos aparecen en ese lado, y entonces la media se extrema pues es sensible a la presencia de casos atípicos: $Mo < Me < \bar{x}$. Si la asimetría es negativa la distribución está sesgada por la izquierda y en ese caso: $\bar{x} < Me < Mo$.



$\bar{x} < Me < Mo$
Sesgo a la izquierda

$Mo < Me < \bar{x}$
Sesgo a la derecha

Para determinar la simetría se puede calcular el coeficiente de asimetría de Pearson que se define como:

$$AP = \frac{\bar{x} - Mo}{s} \qquad \textit{Ecuación 20}$$

Se utiliza poco pues se requiere quela distribución sea unimodal, con forma de campana y que moderada o ligeramente asimétrica.

Alternativamente se calcula el coeficiente de asimetría de Fisher:

$$AF = \gamma_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 \Big/ n}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 \Big/ n-1\right)^{3/2}} \qquad \textit{Ecuación 21}$$

Si el valor del coeficiente es positivo, la distribución será asimétrica positiva, hacia la derecha. Si el coeficiente es igual a cero, la distribución será perfectamente simétrica. Finalmente, si el coeficiente es negativo, la distribución será asimétrica negativa, hacia la izquierda.

En los dos casos de las variables edad y PIB per cápita de la Tabla III.3.8 se observa que la asimetría es positiva, se desvían hacia la derecha las personas mayores y los países ricos, con diferencias de magnitud notables: 0,265 y 2,222, respectivamente.

3.4.2. *Curtosis*

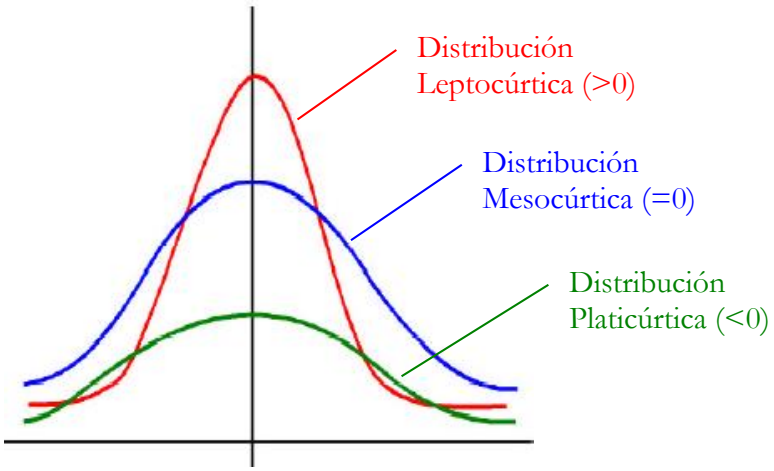
La curtosis es un coeficiente que nos indica el grado de apuntamiento achatamiento que presenta una distribución de datos respecto a la distribución normal estándar, por tanto, solamente tiene sentido condiderar esta medida con distribuciones unimodales y simétricas.

El coeficiente de curtosis de Fisher se calcula mediante:

$$CF = g_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 \Big/ n}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \Big/ n} - 3$$

Ecuación 22

Si el coeficiente es positivo diremos que la distribución es leptocúrtica o apuntalada, indica que las observaciones se concentran más y presentan colas menos largas que las de una distribución normal. Si el coeficiente es igual a cero, diremos que la distribución es mesocúrtica como la distribución normal. Por último, si el coeficiente es negativo, diremos que la distribución es platicúrtica o plana o achatada, indica que las observaciones se agrupan menos y presentan colas más largas.



Las variables edad y PIB per cápita de la Tabla III.3.8 no pueden interpretarse estrictamente dada su falta de simetría. Si lo fueran, en un caso tendríamos una distribución platicúrtica (la edad con el valor -0,790) y en el otro una distribución marcadamente leptocúrtica (el PIB per cápita con el valor 6,677).

4. El análisis exploratorio de datos

El análisis exploratorio de los datos es un procedimiento de análisis descriptivo para resumir la información de una distribución de frecuencias de una variable cuantitativa con un mínimo de pérdida de información detectando la existencia de casos extremos y otras singularidades. Al mismo tiempo realiza pruebas de bondad para distintas

condiciones de comportamiento de la distribución que son necesarias cuando se realizan contrastes inferenciales de hipótesis estadísticas. Sobre este último aspecto nos detendremos en el próximo capítulo.

El Análisis Exploratorio de los Datos es el enfoque que fundamenta este tipo de procedimiento, desarrollado originalmente por John Tukey como alternativa al análisis clásico basado en la media y en la desviación típica. En este caso se propone de medidas de tendencia central a la mediana y la media recortada junto a los cuartiles y el rango intercuartil como medida de dispersión. Se trata de un procedimiento complementario al análisis más tradicional donde los cálculos de resumen están basados en la media, teniendo en cuenta por tanto la totalidad de los casos. Por el contrario la mediana tan sólo utiliza el valor central, descartando todos los demás. Su utilidad es relevante cuando nuestro interés radica en estimar tendencias centrales y nos encontramos con casos extremos -obtenidos por error o por ser atípicos- que desvirtúan el valor de la media como ya hemos destacado con anterioridad.

En un análisis exploratorio descriptivo de hecho contemplamos igualmente los estadísticos o descriptivos que hemos visto hasta ahora de posición, dispersión y forma, y se incorpora una representación gráfica de interés denominada diagrama de caja. Presentaremos seguidamente los resultados que se obtienen de analizar las variables edad y PIB per cápita.

Tabla III.3.11. Descriptivos del análisis exploratorio de las variables P32 (edad) y del PIB per cápita

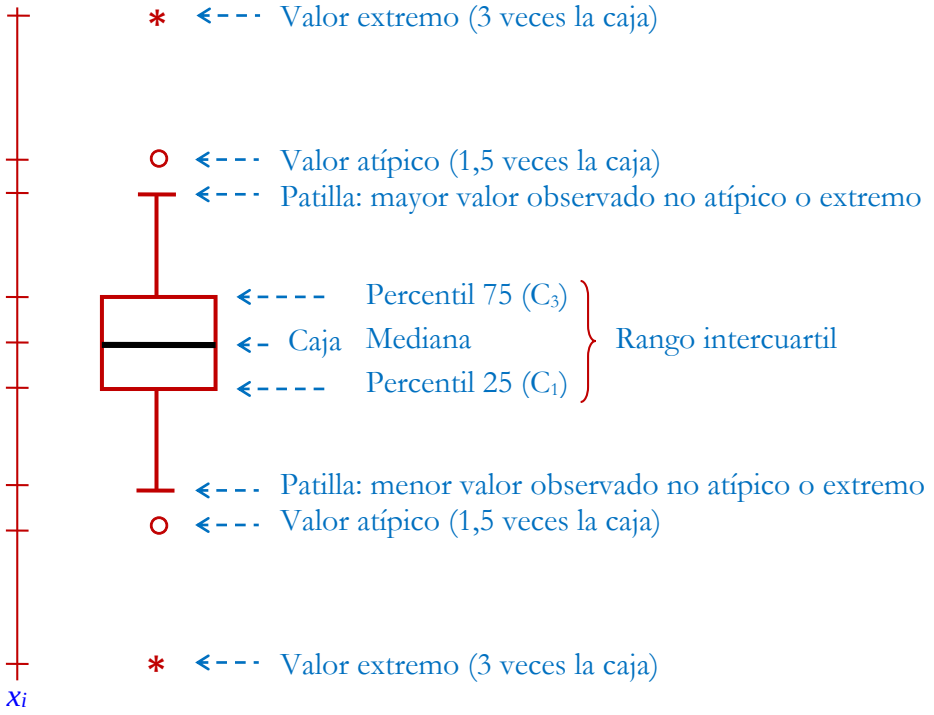
		P32 Edad	PIB per cápita
Media	Estadístico	48,32	16486,51
	Error estándar	0,35	1333,64
95% de intervalo de confianza para la media	Límite inferior	47,63	13855,78
	Límite superior	49,00	19117,25
Media recortada al 5%		47,95	14202,93
Mediana		46,00	10338,97
Varianza		305,86	337934746,57
Desviación estándar		17,49	18383,00
Mínimo		18	443,96
Máximo		94	119029,12
Rango		76	118585,16
Rango intercuartil		28	18396,05
Percentiles	5	22	1135,96
	10	25	1499,93
	25	34	3629,59
	50	46	10338,97
	75	62	22025,63
	90	73	41850,53
	95	79	53003,76
Asimetría	Estadístico	0,265	2,222
	Error estándar	0,049	0,176
Curtosis	Estadístico	-0,790	6,677
	Error estándar	0,098	0,351

En la tabla Tabla III.3.11 vemos el estadístico de la media recortada al 5%, un descriptivo a medio camino entre la media y la mediana, que calcula la media de los casos comprendidos en un intervalo central de la distribución, en este caso excluyendo el 5% de cada lado con el objetivo de eliminar del cálculo los posibles casos extremos o atípicos, obteniendo una media más representativa que la media aritmética.
Si comparamos la media y la media recortada de la variable edad podemos apreciar que no se produce una variación importante (48,32 frente a 47,95) por lo que nos indica que la distribución no se extrema ni tiene casos atípicos que influyan de forma determinante en el cálculo de la media. Si se reduce algo es porque justamente había algunos pocos individuos de mayor edad (sesgo por la derecha) que tiraban de la media hacia ellos; al eliminarlos, la media se desplaza hacia la izquierda al perder la influencia de aquéllos.

En el caso de la variable del PIB per cápita sabemos que la distribución en mucho más asimétrica y que existen algunos casos de países muy rico que difieren de la mayoría del resto de países del mundo. Esta circunstancia se refleja en el cambio del valor de la media aritmética que pasa de 16.486,51 a 14202,93 dólares, una diferencia notable derivada de la influencia de unos pocos países extremadamente ricos.

Del mismo modo que la media la desviación típica se ve afectada por los casos extremos. De forma alternativa existe el estadístico del rango intercuartil que mide la diferencia que hay entre el valor del cuartil superior y el cuartil inferior, permitiendo afirmar que el 50% de los casos de cualquier distribución se encuentran entre ambos cuartiles. Este descriptivo servirá para determinar la presencia de comportamientos extremos en la representación gráfica del diagrama de caja.

El esquema siguiente sintetiza las características de un diagrama de caja:

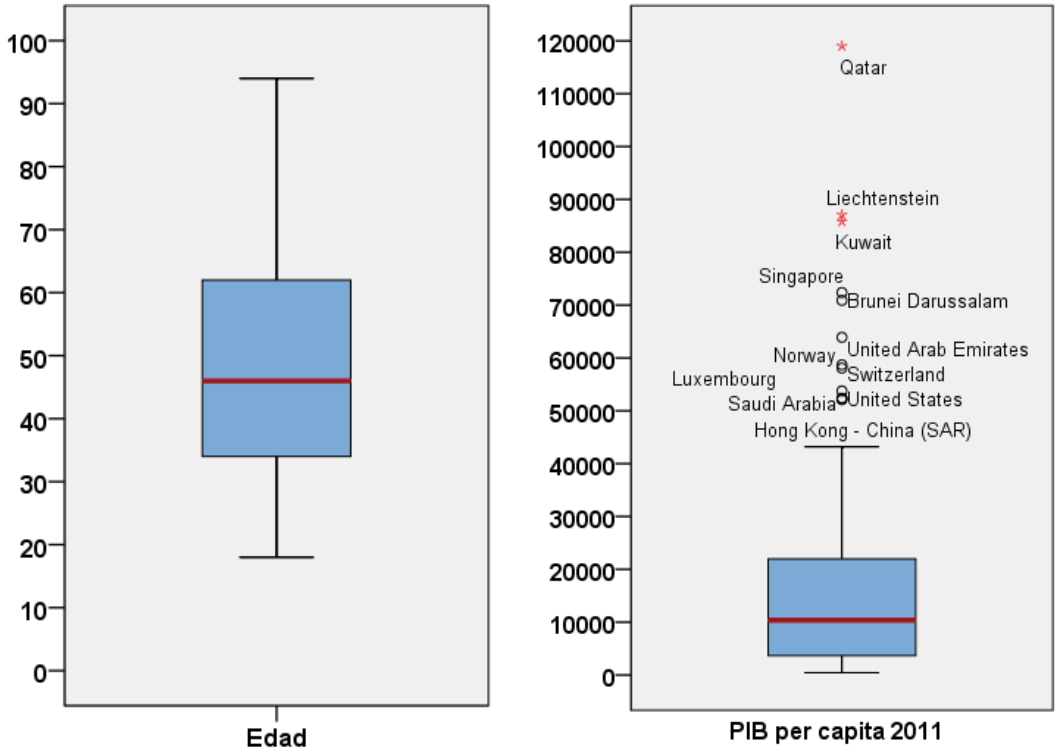


El diagrama de caja (*box plot*) es un gráfico basado en las medidas de posición que ofrece un resumen de la información más relevante de la distribución a partir de la mediana, los percentiles 25 y 75 (primer y tercer cuartiles), y aquellos casos de valores que superan 1,5 veces el rango intercuartil (los atípicos, *outliers*) o 3 veces el rango intercuartil (los extremos o atípicos severos). Esquemáticamente se puede presentar así:

Las patillas (o bigotes) marcan el límite de 1,5 veces el rango intercuartil, es decir, 1,5 veces la longitud de la caja, y a partir del cual consideramos que se trata de un valor atípico en relación al conjunto más centrado de la distribución. Si no hubiera casos atípicos la posición de la patilla puede ser mayor o menor dependiendo del grado de concentración o dispersión. En una situación de elevada concentración la patilla puede llegar confundirse con el extremo de la caja. Los valores atípicos y extremos se suelen presentar con el número de caso (o una etiqueta de identificación) al que corresponden para identificarlo, analizarlo y eventualmente eliminarlo del análisis para que no afecte en los cálculos y en el análisis.

En el Gráfico III.3.15 se representan los diagramas de cajas de la edad y el PIB per cápita. En el primer caso no se observan casos atípicos y en segundo tenemos tanto atípicos como extremos. La mediana en los dos casos no se sitúan en medio de la caja sino que se desplaza hacia abajo (el centro de la caja que implicaría una distribución simétrica), expresando la mayor concentración que existe en los valores bajos de las variables (jóvenes y pobres, respectivamente), sobre todo en el caso del PIB per cápita.

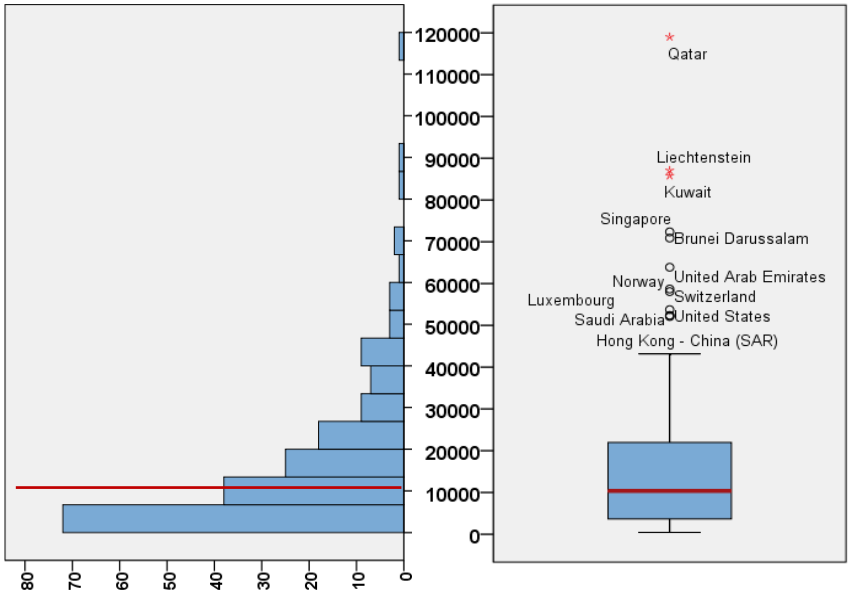
Gráfico III.3.15. Diagramas de caja de las variables P32 (edad) y PIB per cápita



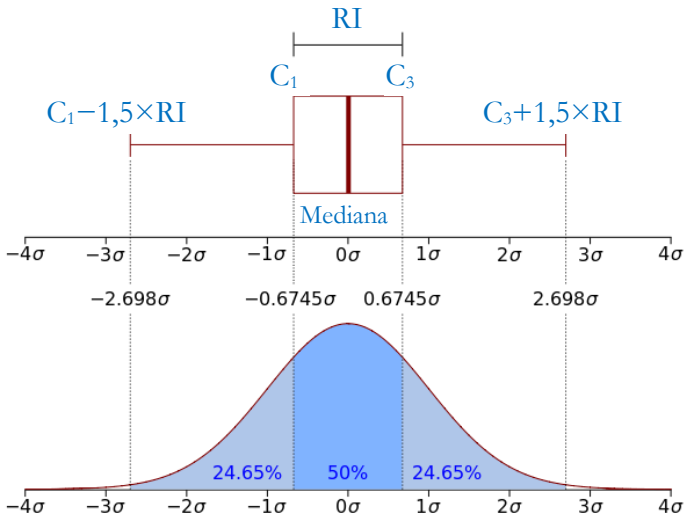
Las patillas en el caso de la edad se sitúan en los valores mínimo y máximo de la distribución, y ningún caso llega a ser 1,5 veces el rango intercuartil ($1{,}5 \times RI = 1{,}5 \times 28$

= 42 *años*): por abajo $1{,}5\times RI$ a partir de la caja sería $C_1-1{,}5\times RI$, es decir, 34−42, y no hay nadie con edades negativas; por arriba $1{,}5\times RI$ a partir de la caja sería $C_3+1{,}5\times RI$, es decir, 62+42, y no hay nadie de 104 años. Las patillas corresponden por tanto al valor mínimo, 18, y al valor máximo, 94. En el caso del PIB per cápita la patilla superior sí que se sitúa a $1{,}5\times RI$ pues se dan casos de valores atípicos, como Estados Unidos o Noruega, y extremos, como Catar o Liechtenstein.

La distribución del PIB per cápita muestra pues un comportamiento mucho más asimétrico y disperso que la edad, como también habíamos tenido ocasión de ver con los histogramas y las otras medidas de resumen. Son formas perfectamente complementarias de analizar la información. En el gráfico siguiente se puede apreciar la correspondencia entre el diagrama de caja y el histograma para dar cuenta de la distribución:



La simetría perfecta de una distribución se representaría de esta forma:



5. Transformación de los datos

Como vimos en el capítulo anterior diversas son las transformaciones de los datos para realizar el análisis de la información estadística. Fundamentalmente vimos dos tipos de transformaciones: la recodificación de valores de una variables, ya sea para cambiar los valores de alfanuméricos a numéricos, para cambiar los códigos iniciales o para agruparlos, y la generación de nuevas variables como resultado de la combinación de varías de ellas entre sí para crear tipologías, una tasa o un cálculo funcional. Aquí nos centraremos en estas últimas para dar cuenta sobre todo de un tipo de transformación de gran utilidad en el análisis estadístico como es la denominada tipificación o estandarización de las variables. Otras transformaciones funcionales podría ser el cálculo de un indicador como la densidad de población, resultado de dividir la población entre la extensión de un territorio, o la renta per cápita cuando dividimos la renta entre la población.

5.1. Transformaciones funcionales básicas

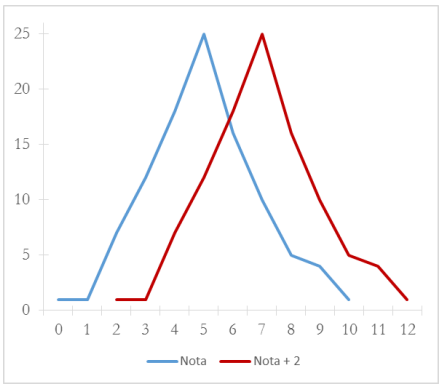
Las funciones básicas que se pueden aplicar a una o más variables cuantitativas son: sumar o restar constantes, sumar o restar variables, y multiplicar por constantes. La tipificación de una variable que veremos más tarde será una combinación de estas transformaciones.

Cuando sumamos o restamos a cada valor de una variable un valor constante K el resultado es una traslación de la distribución de la variable. Si la constante es positiva, la translación es un desplazamiento de la distribución hacia la derecha si es negativa la translación es hacia la izquierda. Con las translación todos los casos se desplazan por igual conservando la forma de la distribución. Por tanto, las características de posición se verán afectadas por la constante mientras que las características de dispersión, asimetría y curtosis de la nueva variable serán las mismas que las de la variable inicial.

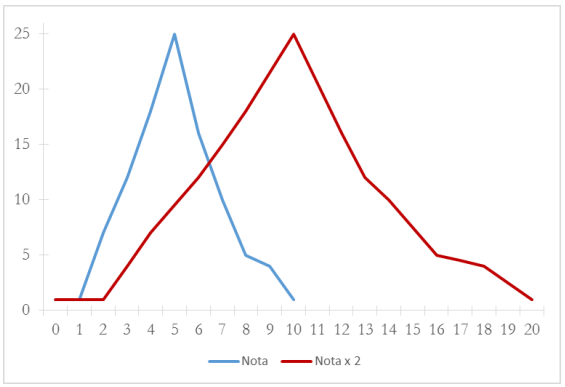
Cuando multiplicamos cada valor de una variable por un mismo valor constante K el resultado es un cambio de escala (o una homotecia) de un factor igual al valor de esa constante. Si el valor absoluto del factor es mayor que 1 ($|K|>1$), la dispersión de los datos aumenta: la escala se amplía. Si el valor absoluto es 1 ($|K|=1$), la dispersión no cambia y la escala se queda igual. Finalmente, si el valor absoluto es menor que 1 ($|K|<1$), la dispersión disminuye: la escala se reduce. Además, si la constante es negativa, las observaciones cambian de lado del eje de abscisas. Por lo tanto, las características de posición se verán afectadas por el factor de escala de magnitud $|K|$ al igual que las características dispersión en una magnitud también $|K|$, aunque en el caso de la varianza, por ser un cálculo cuadrático, se verá por K^2. En este caso la forma de la distribución tampoco varía con un cambio de escala.

En los polígonos de frecuencias que se representan continuación podemos observar los efectos de las dos operaciones de transformación. A partir de una distribución inicial de las notas de un grupo de 100 alumnos/as con media 5 y desviación típica 1,9, en el primer caso se suman dos puntos a cada nota, se suma una constante $K=2$. El resultado es una nueva distribución idéntica a la anterior en forma y desplazada dos

unidades hacia la derecha. Ahora la media es 7 y la desviación típica permanece inalterada en 1,9. En el segundo caso la notas se multiplican por 2, por un factor de escala $K=2$. Como resultado la distribución se amplía cambiando la escala de los valores: las notas que iban de 0 a 10 ahora van de 0 a 20, por lo que la media pasa de 5 a 10 y la desviación de 1,9 a 3,8.



(a) Traslación



(b) Cambio de escala

Una transformación que combina una traslación y una homotecia implica que se suma una constante (a) y se multiplica por una constante (b), se denomina transformación afín, y es el caso de una función lineal: $y=a+bx$.

Otras transformaciones funcionales de las variables serían las que se obtienen con la función logarítmica, la función exponencial, la potencial o la polinómica, que son de utilidad en el tratamiento de las variables en diversas técnicas de análisis de datos como en la regresión lineal o en el análisis de datos longitudinales.

5.2. Tipificación de las variables

Las variables cuantitativas originales que se obtienen en cualquier investigación o fuente de datos vienen caracterizadas por una unidad de medida propia de la característica que expresa cada una de ellas. Con cada variable podemos calcular una media y una desviación para resumir las características de su distribución. En estas condiciones cabe plantearse dos cuestiones:

- ¿Hasta qué punto el valor o la puntuación de esta persona difiere del resto? ¿es una persona muy atípica o no? Para valorar su atipicidad, su diferencia con respecto a las demás personas, podemos calcular la diferencia entre su valor y la media de la distribución de todas las personas, y obtendremos un valor absoluto de su variación con respecto de la media. De esta forma centrar la distribución como vimos a la hora de explicar la varianza.

- Por otro lado, si queremos comparar ambas variables entre sí nos encontramos con el problema de que los valores están expresados en unidades distintas. Para posibilitar la comparación, además de calcular la diferencia con respecto de la media, podemos dividir el resultado por la desviación típica correspondiente. De esta forma conseguimos convertir los valores originales en valores centrados

expresados en unidades de desviación, consiguiendo así estandarizar o tipificar la variable. Estos valores también se denominan puntuaciones típicas. De esta forma se reducir o se cambia de escala la distribución.

La tipificación o estandarización de una variable x crea una nueva variable z, expresada en unidades de desviación, donde cada valor estandarizado indica el número de unidades de desviación por encima o por debajo de la media de la variable. Su fórmula de cálculo es:

$$z_i = \frac{x_i - \bar{x}}{s} \qquad \textit{Ecuación 23}$$

Tipificar o estandarizar las variables es una operación fundamental en estadística y se utiliza en múltiples procedimientos de análisis estadístico y de tratamiento de la información. Su interés reside en dos propiedades principales:

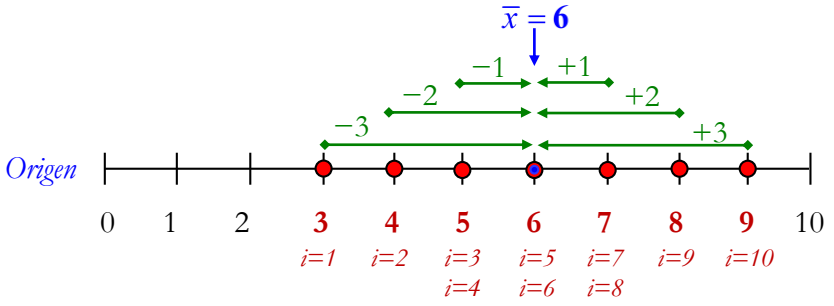
- La variable estandarizada z_i se caracteriza por tener media 0 y desviación típica igual a 1, siendo la unidad de medida comparable y expresada en términos de unidades de desviación.
- Permite relacionar las puntuaciones típicas z_i con la distribución normal.

Veamos un ejemplo de cálculo a partir de los datos de las notas que vimos al comentar la media y la varianza. En la tabla siguiente se realiza la tipificación para cada individuo a partir de una media de 6 y una desviación de 1,83:

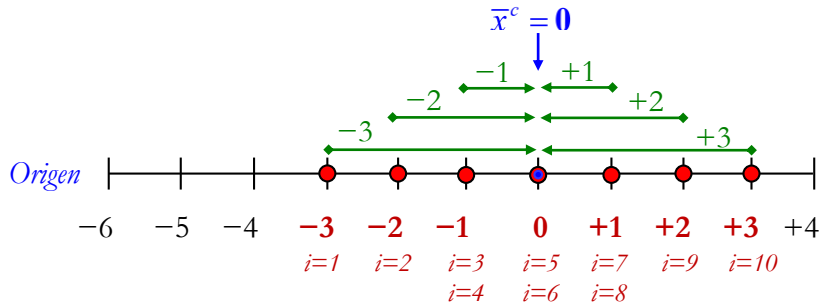
		Datos centrados	*Datos tipificados*
i	x_i	$x_i^c = x_i - \bar{x}$	$z_i = \frac{x_i - \bar{x}}{s}$
1	3	-3	−1,64
2	4	-2	−1,10
3	5	-1	−0,55
4	5	-1	−0,55
5	6	0	0,00
6	6	0	0,00
7	7	1	0,55
8	7	1	0,55
9	8	2	1,10
10	9	3	1,64
Total	60	0	0,00

Representábamos la información de esta manera cuando centrábamos los datos a partir de la información original:

Datos originales

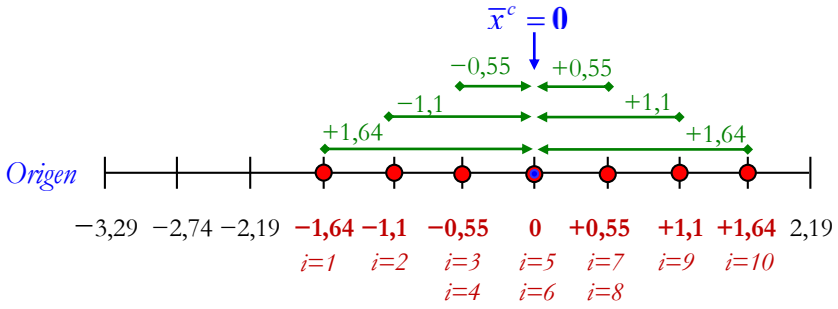


Datos centrados



Ahora los datos tipificados implican una representación similar pero con cambio de escala:

Datos tipificados



Obtenemos así unos valores nuevos, que representan a los mismos originales, pero se expresan en términos el número de desviaciones típicas que nos desviamos de la media. Para el caso de 0,55 unidades de desviación en la nueva escala, si lo multiplicamos por el valor de una desviación típica, 1,83, el resultado es el valor original de l con los datos centrados[17], y de 7 si le sumamos a continuación la media. Obtenemos así datos estandarizados que podemos comprar con cualquier otra variable que esté expresada también esté expresada en términos de unidades de desviación típica.

[17] Es la misma situación que se produce cuando cambiamos la moneda, el tipo de cambio sería el valor de la desviación típica que nos sitúa en una nueva escala de valoración de los precios.

6. Bibliografía

Agresti, A.; Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Upper Saddle RIver (New Jersey): Pearson Prentice Hall. 4ª edición.

Bardina, X.; Farré, M.; López-Roldán, P. (2005). *Estadística: un curs introductori per a estudiants de ciències socials i humanes. Volum 2: Descriptiva i exploratòria bivariant*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona. Col·lecció Materials, 166.

Blalock, H. M. (1978). *Estadística Social*. México: Fondo de Cultura Económica.

Bryman, A., Duncan, C. (1990). *Quantitative data analysis for social scientists*. London: Routledge.

Dalgaard, P. (2008). *Introductory Statistics with R*. New York: Springer.

Domènech, J. M. (1982). *Bioestadística. Métodos estadísticos para investigadores*. Barcelona: Herder. 4ª. Edición.

Domínguez, M.; Simó, M. (2003). *Tècniques d'Investigació Social Quantitatives*. Edicions de la Universitat de Barcelona. Col·lecció Metodologia, 13.

Everitt, B. S.; Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. New York: Cambridge University Press. Fourth Edition.

Finkelstein, M. O.; Levin, B. (2001). *Statistics for Social Science and Public Policy*. New York: Springer. 2ª edición.

Farré, M. (2005). *Estadística: un curs introductori per a estudiants de ciències socials i humanes. Volum 1: Descriptiva i exploratòria univariant*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona. Col·lecció Materials, 162.

Farré, M.; Bardina, X. (2009). *Estadística descriptiva*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona. Manuals, 54.

Feldman, J.; Lagneau, G.; Matalon, B. (1991). *Moyenne, milieu, centre. Histoires et usages*. Paris: École des Hautes Études en Sciences Sociales.

García Ferrando, M.; Ibañez, J.; Alvira, F. (1986). *El análisis de la realidad social. Métodos y técnicas de investigación*. Madrid: Alianza.

García Ferrando, M. (1994) *Socioestadística. Introducción a la estadística en sociología*. 2a edició rev. i amp. Madrid: Alianza. Alianza Universidad Textos, 96.

Guisan de González, C.; Vaamonde Liste, A,; Barreiro Felpeto, A. (2011). *Tratamiento de datos con R, STATISTICA y SPSS*. Madrid: Díaz Santos.

Hopkins, K. D.; Hopkins, B. R.; Glass, G. V. (1997). *Estadística Básica para las ciencias sociales y del comportamiento*. 3ª. edición. Naucalpan de Juárez: Prentice-Hall Hispanoamericana.

IBM Corporation (2015b). *IBM SPSS Statistics Base 22*. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/es/client/Manuals/IBM_SPSS_Statistics_Base.pdf.

IBM Corporation (2015c). *Guía breve de IBM SPSS Statistics 22*. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/es/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf.

Landau, S.; Everitt, B. S. (2004). *A Handbook of Statistical Analyses using SPSS*. Boca Raton: Chapman & Hall/CRC.

Marqués, J. P. (2007). *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*. Berlin: Springer. 2ª edición.

Moncho, J. (2015). *Estadística aplicada a las ciencias de la salud*. Barcelona: Elsevier.

Morgan, G. A. et al. (2004). *SPSS for Introductory Statistics. Use and Interpretation*. Mahwah (New Jersey): Lawrence Erlbaum.

Pardo, A. (2002). *SPSS 11 Guía para el análisis de datos*. Madrid: McGraw-Hill.

Pardo, A.; Ruíz, M. A. (2005). *Análisis de datos con SPSS 13 Base*. Madrid: McGraw-Hill.

Peña, D. (1997). *Introducción a la Estadística para las Ciencias Sociales*. Madrid: McGraw Hill.

Peró, M. et al. (2012). *Estadística aplicada a las ciencias sociales mediante R y R-Commander*. Madrid: Ibergarceta Publicaciones.

Ritchey, F. J. (2008). *Estadística para las ciencias sociales*. Madrid: McGraw Hill.

Sánchez Carrión, J. J. (1999). *Manual de análisis estadístico de los datos*. Madrid: Alianza. Manuales 055.

Visauta Vinacua, B. (2002). *Análisis estadístico con SPSS 11.0 para Windows. Volumen I. Estadística básica*. 2a. Edició. Madrid: McGraw-Hill.

Webster, A. L. (2000). *Estadística Aplicada a los negocios y la economía*. Santa Fe de Bogotá: McGraw-Hill.

Wonacott, Th. H.; Wonacott, R. J. (1979). *Fundamentos de Estadística para Administración y Economía*. México: Limusa.