Diseño de portada:
Edita:
ISBN:
Depósito Legal:
Imprime:

INICIACIÓN AL ANÁLISIS DE DATOS CUANTITATIVOS EN EDUCACIÓN. TEORÍA Y PRÁCTICA MEDIANTE SPSS DEL ANÁLISIS DESCRIPTIVO BÁSICO

AUTORES:

Clemente Rodríguez Sabiote
Miguel Ángel Gallardo Vigil
Teresa Pozo Lorente
José Gutiérrez Pérez

INDICE

PRIMERA PARTE: TEORÍA

1. Ideas previas sobre la organización de datos en la investigación educativa

Organizar los datos recogidos en una investigación educativa, supone el primer paso para poder llevar a cabo interpretaciones de los mismos y formular conclusiones. Los procedimientos sobre el uso de la organización de los datos parten de una idea elemental: tomar decisiones de síntesis, agrupamiento y simplificación para poder formular conclusiones. A este propósito, sirven los procedimientos de organización de datos de forma creativa y sencilla cuando existen pocos datos.

Ejemplo: El profesor de 2º de E.S.O. de Tecnología de un centro educativo ha realizado un ejercicio con los 14 alumnos de nuevo ingreso en el centro. Las puntuaciones que han obtenido se presentan a continuación de dos formas distintas:

Presentación A

Sujetos	Puntuación
1	45
2	44
3	32
4	49
5	44
6	50
7	49
8	44
9	50
10	50
11	45
12	32
13	32
14	49

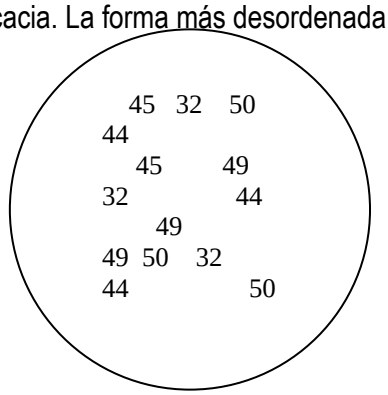
Presentación B

Puntuación 32
Sujeto 3 / Sujeto 12 / Sujeto 13
Puntuación 44
Sujeto 2 / Sujeto 5 / Sujeto 8
Puntuación 45
Sujeto 1 / Sujeto 11
Puntuación 49
Sujeto 4 / Sujeto 7 / Sujeto 14
Puntuación 50
Sujeto 6 / Sujeto 9 / Sujeto 11

1. Observa la presentación A, ¿puedes sacar alguna conclusión sobre las puntuaciones de los alumnos?
2. Ahora observa la presentación B, ¿se te ocurre alguna conclusión a golpe de vista?

En este ejemplo, los datos obtenidos son muy pocos, por lo que formular conclusiones es fácil, aún cuando no hayan sido ordenados por ningún procedimiento. Pero si tenemos

grandes cantidades de datos es necesario hacer uso de sistemas convencionales de probada eficacia. La forma más desordenada de presentar estos datos sería:



Los sistemas convencionales de organización de datos encierran acuerdos, normas y convenciones sobre sistemas exitosos para transmitir información con los datos. Veamos varios procedimientos para organizar estos datos siguiendo reglas, normas y procedimientos más o menos convencionales:

a) Por orden creciente: 32, 32, 32, 44, 44, 44, 45, 45, 49, 49, 49, 50, 50, 50

b) Por orden decreciente: 50, 50, 50, 49, 49, 49, 45, 45, , 44, 44, 44, 32, 32, 32

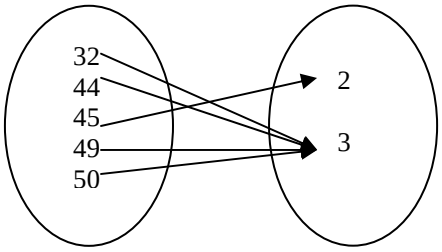
c) Por agrupaciones según repeticiones: tres 32, tres 44, dos 45, tres 49, tres 50

d) Por agrupaciones según las veces que aparecen:

Dos veces: 45

Tres veces: 32, 44, 45, 49, 50

e) Mediante diagramas de Ven



f) Se te ocurre algún procedimiento creativo para organizar estos datos. ¡Inténtalo!

Los sistemas de organización y representación de datos más empleados son las tablas, los diagramas, los sistemas de representación estadística convencionales, así como determinados procedimientos numéricos como el análisis exploratorio de datos en diagramas de tallo y hoja.

Del dato bruto a la puntuación transformada y codificada para poder ser interpretada y sacar conclusiones hay diferentes opciones según el tipo de datos, la cantidad de los mismos y la finalidad del análisis, una síntesis de las más usuales son: los datos brutos ordenados, los datos organizados en tablas de frecuencias, los datos organizados en intervalos, los datos representados visualmente mediante gráficos creativos, los datos organizados en diagramas de

tallo y hojas (análisis exploratorio de datos); los datos transformados mediante procedimientos matemáticos sencillos (frecuencias, porcentajes, proporciones) o estadísticos de síntesis más elaborados (media, mediana, moda y medidas de variación).

Para poder comprender mejor cada una de estas formas de presentar la información lo haremos partiendo del siguiente ejemplo: La Biblioteca de la Facultad de Educación está realizando un estudio sobre el número de libros que prestan al alumnado durante el mes de marzo. El total de alumnos del estudio ha sido 108 correspondientes a primer curso de la Diplomatura de Maestro especialista en Educación Infantil.

La información aparece en la siguiente tabla:

1	2	1	3	4	5	1	1	2	3	4	5	6	3	2	1	1	2
2	3	4	5	5	5	4	4	6	3	2	2	3	3	3	3	4	3
3	2	2	1	1	6	6	5	5	5	5	3	2	3	4	5	2	1
6	5	4	5	6	4	3	5	5	6	6	3	2	2	1	1	3	4
4	6	6	6	1	1	2	2	2	3	4	5	6	5	4	3	2	1
1	2	2	3	4	5	4	5	5	6	4	5	5	5	6	5	5	2

a) Los datos brutos ordenados. Como su nombre indica se centra en presentar todos los datos obtenidos ordenados. El primer paso en la tarea de análisis se centra en realizar una ordenación de los mismos

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
5	5	5	5	6	6	6	6	6	6	6	6	6	6	6	6	6	6

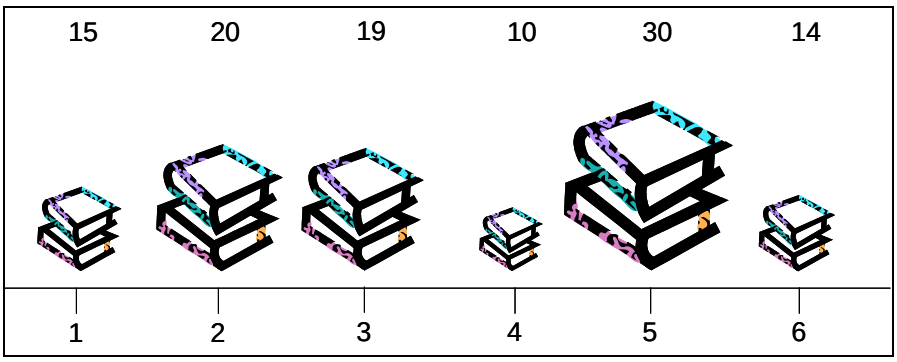
b) Los datos organizados en tablas de frecuencias. Podemos observar que la información, aún estando ordenada, tiene el inconveniente de la extensión de la misma. Para ello podemos utilizar las tablas de frecuencias en las que presentamos de forma ordenada las puntuaciones que hemos obtenido y a su derecha el número de veces que aparece, es decir, su frecuencia.

Puntuación	Frecuencia
1	15
2	20
3	19
4	10
5	30
6	14
∑	108

c) Los datos organizados en intervalos. En nuestro ejemplo partimos de un total de 108 alumnos con puntuaciones que oscilan entre 1 y 6 (número de libros), pero a podemos encontrarnos con un mayor número de sujetos y de valores. Para ello podemos organizar los datos en intervalos.

Intervalo	Frecuencia
1-2	35
3-4	29
5-6	44
∑	108

d) Los datos representados visualmente mediante gráficos creativos. La información presentada a través de gráficos nos ofrece una visión general de los datos, que con un simple vistazo podemos interpretar.



e) Los datos organizados en diagramas de tallo y hojas (análisis exploratorio de datos). El desarrollo del análisis exploratorio de datos a partir de las propuestas de Tuckey (1977) ha supuesto un importante revulsivo en el uso de estrategias de organización de datos. Los diagramas de tallo y hojas o las representaciones orientadas por los principios de la estadística visual hacen posible que el

destinatario de la información de la investigación pueda interpretar y entender sin ser experto en complejas estrategias estadístico- matemáticas.

```
VAR00001 Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    15,00        1 .  000000000000000
    20,00        2 .  00000000000000000000
    19,00        3 .  0000000000000000000
    10,00        4 .  0000000000
    30,00        5 .  000000000000000000000000000000
    14,00        6 .  00000000000000

 Stem width:      1,00
 Each leaf:       1 case(s)
```

f) Los datos transformados mediante procedimientos matemáticos sencillos (frecuencias, porcentajes, proporciones) o estadísticos de síntesis más elaborados (media, mediana, moda y medidas de variación). Otra forma de presentar la información es mediante distintos procedimientos matemáticos que nos ayuden a comprender y poder interpretar mejor los datos recogidos: estadísticos de incidencias, de tendencia central, de variabilidad…

VAR00001

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado	Proporciones
Válidos	1,00	15	13,9	13,9	13,9	0.13
	2,00	20	18,5	18,5	32,4	0.18
	3,00	19	17,6	17,6	50,0	0.17
	4,00	10	9,3	9,3	59,3	0.09
	5,00	30	27,8	27,8	87,0	0.27
	6,00	14	13,0	13,0	100,0	0.12
	Total	108	100,0	100,0		1

Estadísticos

VAR00001

N	Válidos	108
	Perdidos	0
Media		3,5741
Mediana		3,5000
Moda		5,00
Desv. típ.		1,67557
Varianza		2,808

1.1. Análisis exploratorio de datos

Este tipo de análisis consiste en examinar los datos antes de comenzar con la aplicación de cualquier tipo de técnica estadística. Este tipo de análisis proporciona técnicas sencillas para organizar y preparar los datos, detectar fallos en el diseño y su recogida, tratamiento y evaluación de datos ausentes, identificación de casos atípicos.

Para realizar un análisis exploratorio de datos conviene seguir las siguientes etapas (Salvador y Gargallo, 2003):

1) Preparar los datos para hacerlos accesibles a cualquier técnica estadística.

2) Realizar un examen gráfico de la naturaleza de las variables individuales a analizar y un análisis descriptivo numérico que permita cuantificar algunos aspectos gráficos de los datos.

3) Realizar un examen gráfico de las relaciones entre las variables analizadas y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre ellas.

4) Evaluar, si fuera necesario, algunos supuestos básicos subyacentes a muchas técnicas estadísticas como, por ejemplo, la normalidad, linealidad y homocedasticidad.

5) Identificar los posibles casos atípicos (outliers) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.

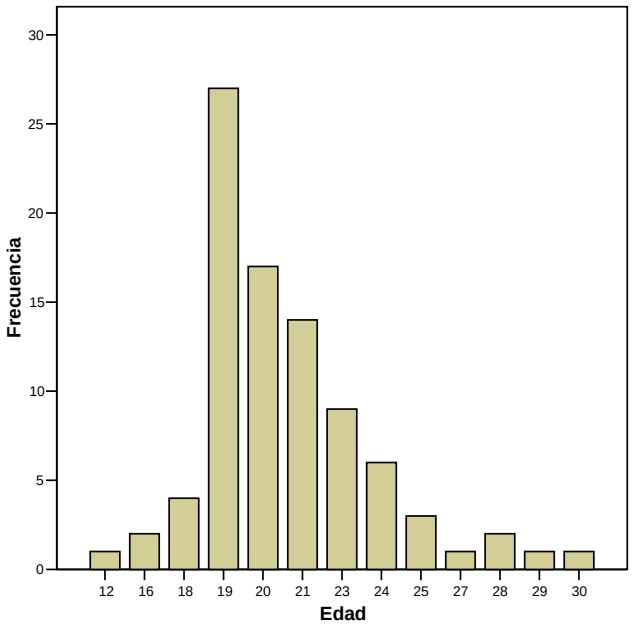
6) Evaluar, si fuera necesario, el impacto potencial que pueden tener los datos ausentes (missing) sobre la representatividad de los datos analizados.

Ejemplo: Se ha realizado una encuesta sobre el uso de las nuevas tecnologías en la docencia práctica del profesorado universitario. A continuación se presenta los datos correspondientes a la variable edad:

19	20	21	19	20	23	23	24	20	19	18	19	21	19	25	30	29	23
12	16	21	25	27	28	21	21	19	19	20	20	23	24	20	19	18	21
23	23	24	20	19	18	19	21	19	25	19	18	19	21	19	19	20	19
28	21	21	19	19	20	20	23	24	20	19	16	20	21	19	19	20	23
24	20	19	21	19	19	20	21	21	19	24	20	19	23	21	19	19	20

Si realizamos una primera exploración de los datos podremos comprobar que alguna de la información que hemos recogido no es válida. Así pues, si la utilizamos para nuestros análisis

no obtendremos unos resultados válidos y fiables. El siguiente gráfico nos ofrece una visión general de nuestros datos:



Estadísticos descriptivos

	N	Media	Desv. típ.
Edad	88	20,81	2,848
N válido (según lista)	88		

Podemos observar que la media de edad de nuestros encuestados es de 20,21 años, pero realmente este valor no es totalmente cierto, ya que encontramos tres puntuaciones que no son válidas. ¿Sabes cuales son?

Efectivamente, tenemos una puntuación de 12 años y dos puntuaciones de 16 años. Estos valores no son válidos, ya que la encuesta está destinada a alumnos universitarios y estos tienen edades superiores o iguales a 18, por lo que cualquier análisis que realicemos con estos datos no nos dará puntuaciones válidas. Si tomamos los valores válidos los resultados serían los siguientes:

Estadísticos descriptivos

	N	Media	Desv. típ.
Edad	85	21,02	2,623
N válido (según lista)	85		

1.2. Algunos ejemplos de presentación de datos de investigaciones reales

A continuación presentamos información obtenida del estudio "Jóvenes y relaciones grupales. Dinámica relacional para los tiempos de trabajo y de ocio (FAD)"[1]

Tabla 6.8. Actividades realizadas con el grupo de amigos durante el fin de semana y entre semana (porcentajes) (respuesta múltiple)

ACTIVIDADES	FIN DE SEMANA	ENTRE SEMANA
Chatear	3.1	7.3
Internet	1.4	2.5
Oír la radio	1.3	4.9
Leer libros, revistas...	0.6	2.1
Estudiar	2.5	13.1
Escuchar cintas, CDs...	4.1	10.3
Ver la televisión	3.6	7.2
Videojuegos	3.2	4.8
Visitar museos, exposiciones	0.5	0.3
Ir a salas de juegos	1.5	0.9
Salir con los amigos sin más	21.9	17.4
Viajar, hacer excursiones	2.4	0.7
Hacer deporte	7.6	14.7
Ir a discotecas, bares...	26.8	2.2
Ir al cine, al teatro...	14.5	4.2
Colaborar con ONGs	0.6	0.8
Ninguna, no nos vemos	0.3	6.0

Otra forma de presentar la información puedes ser como la que a continuación presentamos. Esta información está extraía del estudio "Jóvenes, relaciones familiares y tecnología de la información y las comunicaciones"[2].

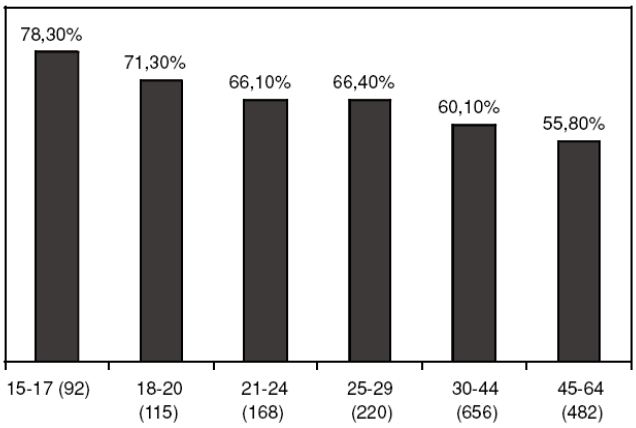
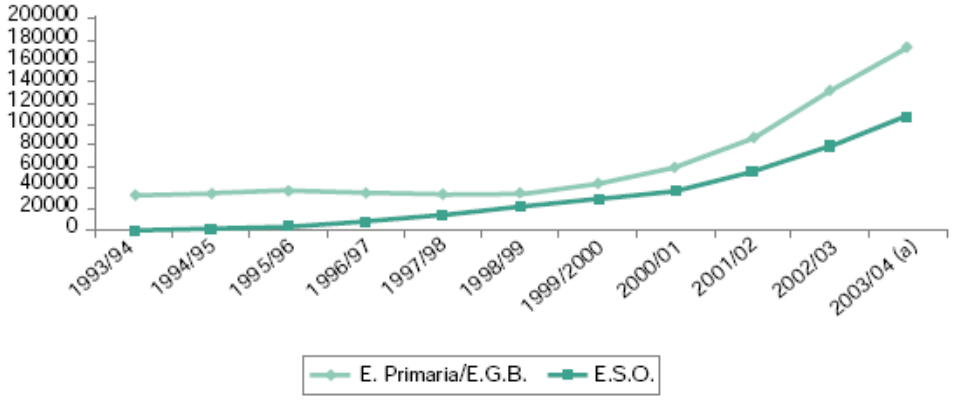


GRÁFICO 6.8
De acuerdo con la frase: *"Con Internet se generan amistades"*, según la edad

[1] Información obtenida del Instituto de la Juventud en la dirección: http://www.injuve.mtas.es/injuve/contenidos.item.action?id=1071722614&menuId= (consultado el 28 de junio de 2006).
[2] Información obtenida del Instituto de la Juventud: http://www.injuve.mtas.es/injuve/contenidos.item.action?id=2062358036&menuId=572069434 (Consultado el 28 de junio de 2006).

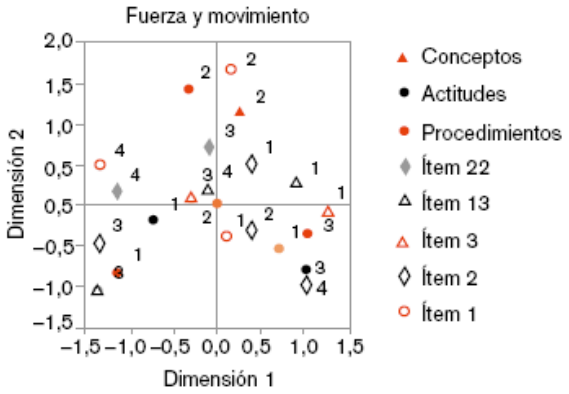
A continuación presentamos un gráfico extraído del estudio sobre "la atención del alumnado inmigrante en el sistema educativo de España"[3].

GRÁFICO 1. EVOLUCIÓN DE LA PRESENCIA DEL ALUMNADO EXTRANJERO MATRICULADO EN EDUCACIÓN PRIMARIA/E.G.B Y E.S.O. AÑOS 1993-2004.



Finalmente, presentamos un gráfico extraído del "Estudio de la influencia de un entorno de simulación por ordenador en el aprendizaje por investigación de la Física en el Bachillerato"[4].

FIGURA 72. REPRESENTACIÓN GRÁFICA DE LAS CUANTIFICACIONES DE LAS CATEGORÍAS DE RESPUESTAS DEL PRE-TEST ASOCIADAS AL DOMINIO CONCEPTUAL DE "FUERZA Y MOVIMIENTO", ASÍ COMO DE LOS DISTINTOS NIVELES DE CONOCIMIENTO CIENTÍFICO.



[3] Información obtenida del Ministerio de Educación y Ciencia: http://www.mec.es/cide/espanol/publicaciones/colecciones/investigacion/col168/col168pc.pdf (Consultado el 30 de junio de 2006).

[4] Información obtenida del Ministerio de Educación y Ciencia: http://www.mec.es/cide/espanol/publicaciones/colecciones/investigacion/col167/col167pc.pdf (Consultado el día 29 de junio de 2006).

2. Nociones básicas sobre análisis descriptivo clásico

2.1. Distribución de frecuencias

Las frecuencias son las medidas que, junto a los porcentajes y proporciones, más se utilizan en el apartado de análisis de datos. Son, desde luego, estadísticos poco complejos, pero que debidamente utilizados e interpretados pueden aportar interesante información a los hallazgos que del estudio desarrollado se derivan.

En realidad, todos sabemos cual es la frecuencia de un determinado valor porque todos podemos llegar a determinar las veces que éste repite. Por ejemplo, supongamos que lanzado un dado 6 veces en 3 ocasiones ha salido 5, en 2 ocasiones el valor 3 y en 1 el valor 2. Con estos precedentes podemos afirmar que las frecuencias de los valores del dado que se han generado son:

Valor del dado	Frecuencia o veces que se repite
2	1
3	2
5	3

Transformar dichas frecuencias en porcentajes y después en proporciones o viceversa es, por tanto, un misión muy fácil, ya que bastaría con, por ejemplo, dividir 1/6 X 100 para el caso de la primera frecuencia y así sucesivamente.

Nos obstante, en la estadística descriptiva clásica se contemplan, además, otras serie de frecuencias fuertemente emparentadas con la anterior. A continuación en la siguiente tabla mostramos dichas frecuencias así como su definición operativa.

CONCEPTOS CLAVES

☑ FRECUENCIA ABSOLUTA: Número de veces que se repite un valor (xi). Se simboliza (fi).

☑ FRECUENCIA RELATIVA: Cociente entre fi de un valor xi y el tamaño de la muestra. Se simboliza (fr), siendo entonces
fr = fi/n.

☑ FRECUENCIA ABSOLUTA ACUMULADA: Suma de cada una de las frecuencias absolutas de cada valor (xi_1, xi_2, xi_3,... xi_n) conformando en cada suma un valor acumulado. Se simboliza como (fia).

☑ FRECUENCIA RELATIVA ACUMULADA: Cociente entre la frecuencia absoluta acumulada de un valor (xi) y el tamaño muestral. Se simboliza como (fra), siendo entonces fra = fia/n.

Ejemplo: Imaginad que el número de errores cometidos por un conjunto de niños (20) al leer un párrafo en una prueba de lectura pasada por un/a Licenciado/a en Pedagogía o Psicopedagogía ha sido el siguiente:

2, 1, 0, 3, 2, 2, 3, 1, 1, 0, 1, 2, 1, 2, 0, 2, 4, 2, 3 y 1. Con estos datos esta sería la tabla de distribución de frecuencias que correspondería al ejemplo citado:

xi (nºerrores)	*fi*	*fr*	*fia*	*fra*
4	*1*	*0,05*	*20*	*1*
3	*3*	*0,15*	*19*	*0,95*
2	*7*	*0,35*	*16*	*0,8*
1	*6*	*0,30*	*9*	*0,45*
0	*3*	*0,15*	*3*	*0,15*
	20	*1*		

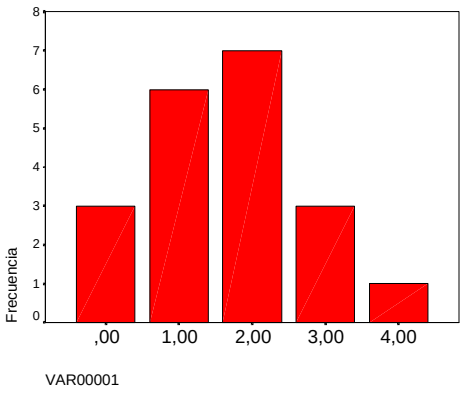
2.2. Representaciones gráficas

A partir de las distribuciones de frecuencias se pueden construir representaciones gráficas. La función de éstas es dar informaciones globales mediante la inspección visual. Siguiendo con el ejemplo anterior mostramos una colección de las representaciones gráficas más usuales.

2.2.1. Representaciones gráficas más frecuentes en el campo de la investigación educativa

2.2.1.1. Diagrama de barras

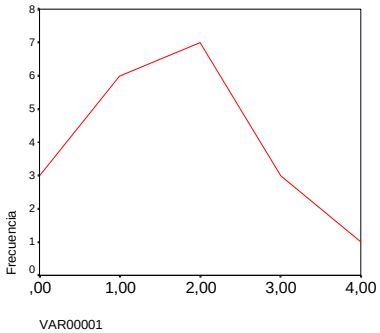
Para construir un diagrama de barras, así como un polígono de frecuencias o histograma debemos contemplar la existencia de dos ejes: ordenadas (y) en vertical y abcisas (x) en horizontal. En el primero, colocaremos los valores obtenidos por las medidas contempladas, mientras que en el segundo las categorías de los mismos. En el caso que explicitamos a continuación las categorías de valores contempladas son: 0, 1, 2, 3 y 4, mientras los valores obtenidos por las frecuencias de cada uno de ellos se representan como 3, 6, 7, 3 y 1 respectivamente. Gráficamente, por tanto, quedaría de la siguiente forma:



2.2.1.2. Pictograma

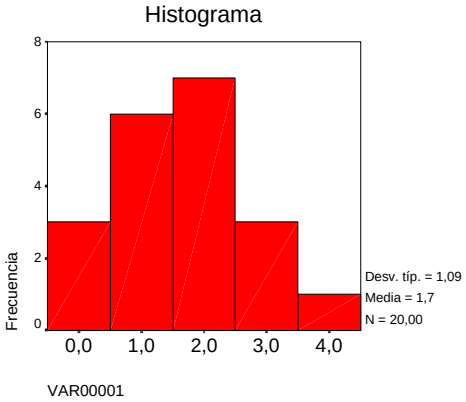
El pictograma es una representación gráfica que utilizada como recurso visual la presencia de algún tipo de dibujo o representación de un elemento. En este caso las diferentes categorías contempladas son agrandadas o empequeñecidas dependiendo de la frecuencia de cada una de ellas. Así por ejemplo las categorías 0 y 3 tendrían el mismo tamaño (frecuencia 3 en ambos casos), mientras 4 sería la más pequeña (fi =1) y las categorías 2 (fi=6) y 3 (fi=7) serían las que tendrían un tamaño mayor.

2.2.1.3. Polígono de frecuencias



2.2.1.4. Histograma

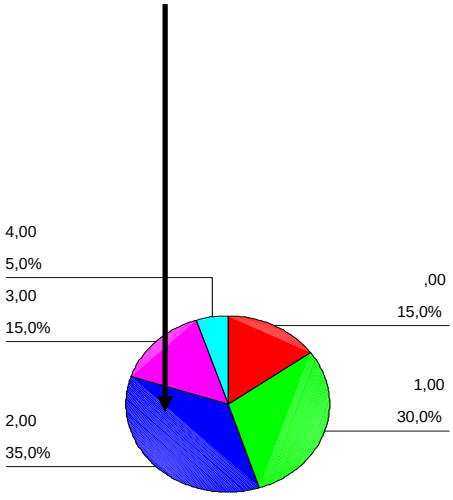
El uso del histograma está indicado cuando la variable a representar se encuentra en una escala de intervalos. En caso contrario se utilizaría el diagrama de barras anteriormente comentado.



2.2.1.5. Diagrama de sectores

El diagrama de sectores es otro de los recursos gráficos que podemos utilizar para la representación de los datos. Al contrario que sus otros compañeros de viaje, este recurso gráfico utiliza un círculo o circunferencia para mostrar la incidencia de los datos. A modo de un pastel o una pizza los datos se representan en porciones que dependen de su incidencia en frecuencias o su equivalente en porcentajes. La amplitud de las porciones no es fortuita y, por ejemplo, el valor 2 (35%) tiene el pedazo o cacho más grande porque es el de mayor frecuencia y, por ende, porcentaje, mientras el caso totalmente contrario es el del valor 4 (5%). En realidad, la representación angular de cada "quesito" es representada, en este caso mediante el programa SPSS, pero cualquier procesador de textos (Word sin ir más lejos) incorpora rutinas de gráficos de estupenda calidad. A modo de ejemplo podemos informarte de cómo el software ha calculado el valor en grados del ángulo del valor 2 (126º):

100% son 360º
35% son X



2.2.1.6. Diagrama de tallo y hojas (stem and leaf)

El diagrama de tallo y hojas es una aportación del estadístico norteamericano John Tukey dentro de la denominada corriente del análisis exploratorio de datos (EDA en su notación anglosajona). Se trata de un gráfico sencillo, intuitivo y muy útil para conocer la forma que adopta la distribución de puntuaciones. En ello se parece a su pariente, la denominada curva normal o campana de Gauss y Laplace. Ambos recursos gráficos sirven para ver donde se producen concentraciones de valores en la distribución de los mismos.

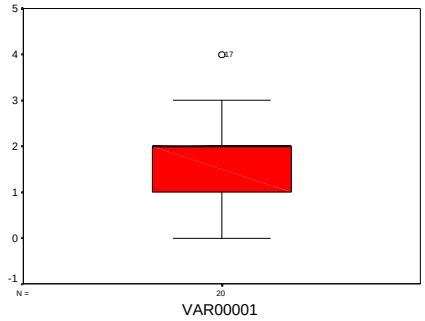
Para elaborar dicho diagrama primero situamos el tallo en la izquierda (Stem) y ahí la categoría de los valores contemplados (0, 1, 2, 3 y 4). En segundo lugar, más a la derecha situamos tantos ceros u hojas como frecuencias haya obtenido dicha categoría. Para interpretar este gráfico debemos inclinar nuestro cuello 90º a la derecha, o mejor mover nuestro cuaderno 90º a la izquierda no vaya ser que nos de una tortícolis. Una vez hecho esto podemos apreciar en qué valores se produce la mayor concentración y si la distribución se asemeja o no a una curva simétrica.

VAR00001 Stem-and-Leaf Plot

Frequency	Stem & Leaf
3	0 . 000
6	1 . 000000
7	2 . 0000000
3	3 . 000
1	4 . 0

2.2.1.7. Diagrama de caja y patillas (box and whiskers)

Mediante este diagrama, al igual que el anterior, podemos averiguar que forma tiene la distribución de nuestros datos. En este caso las dos vallas de los valores 0 y 3 representan los valores mínimo y máximo (aunque esto último no es del todo verdad) de la distribución. Por su parte, los límites superior (cuartil 3) e inferior (cuartil 1) de la caja son los valores 2 y 1 respectivamente. Cuando hemos dicho que el valor máximo no es exactamente el 3 nos referimos a que en realidad el valor máximo y, además, señalado como valor extremo (outlier en su acepción anglosajona) por el programa es el valor 4 obtenido por el sujeto nº 17. En definitiva, la presente representación indicaría que estamos ante una distribución donde la mayoría de valores se acumulan en las categorías 1 y 2.



2.2.2. Algunos errores en la construcción de gráficos

No son pocas las ocasiones en que fortuita o intencionadamente se ilustran gráficos sesgados en informes de investigación. Los errores son de diversos tipos y van desde la manipulación de los ejes de ordenadas, el uso de representaciones tridimensionales hasta la presentación de áreas dimensionales dispares en las representaciones de cada variable. Para tobtener una excelente y extensa información sobre este aspecto el lector interesado puede consultar, entre otras, las obras de Darrell y Geis (1954/1993), Monmonier (2001), Tufte (2003) o Wainer (1997). Nosotros, no obstante, le mostramos algunos sesgos habituales en diagramas de barras y polígonos de frecuencias.

2.2.2.1. La manipulación del eje de ordenadas (o plegamiento de Y)

Dicho sesgo hace referencia a la manipulación que se produce cuando la escala del eje de ordenadas no comienza por el origen (0/1), sino que se establece otro valor con carácter arbitrario) y/o además se transforma la amplitud de los intervalos. Veamos un ejemplo que ilustre el presente sesgo.

Un investigador ha probado un determinado tratamiento para la mejora de la dislexia. Los resultados obtenidos en los grupos experimental (55% superaron una prueba de medición) y control (50% superaron la prueba de medición) apuntan hacia una mejoría discreta del grupo experimental que es posible que ni siquiera sea estadísticamente significativa. No obstante, si el investigador pretende maximizar las propiedades de mejora de dicho tratamiento en vez de presentar el gráfico A, podría presentar el B:

GRÁFICO A

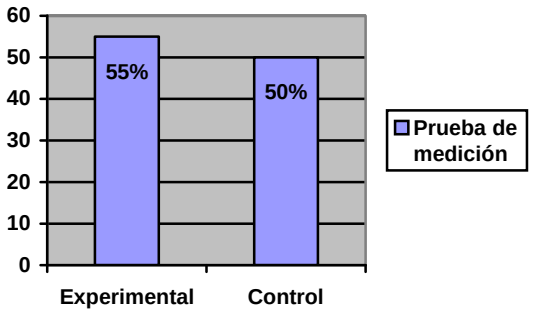
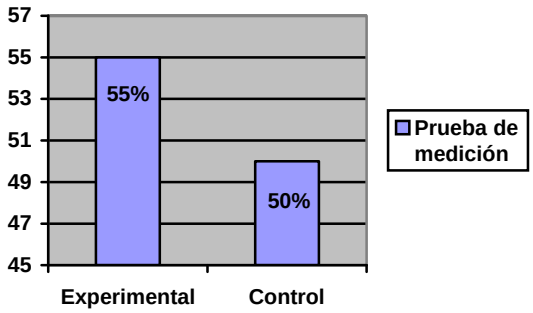


GRÁFICO B



Salta a la vista que si nos quedamos con el gráfico B podríamos pensar, si obviamos la manipulación de la escala del eje de ordenadas, que el tratamiento es realmente eficaz cuando es posible que no lo sea.

Propongamos otro ejemplo: Imaginemos la representación de cinco ítems de una escala tipo Likert (1 a 5) sobre el desarrollo docente en un polígono de frecuencias.

GRÁFICO A

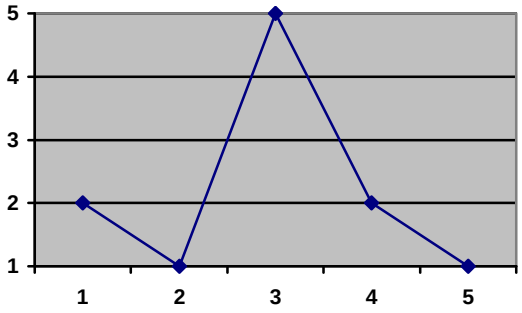
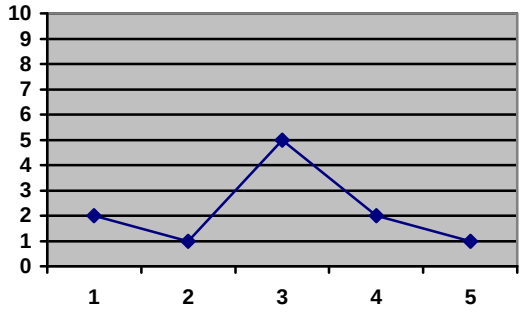


GRÁFICO B



Puede apreciarse como la manipulación del eje Y (ordenadas) sin cambiar la amplitud del intervalo ha generado la mayor o menor pendiente en el patrón de crestas y valles que se dibuja en el gráfico del polígono de frecuencias. Ahora imaginemos que el ítem 3, el que obtiene una puntuación de 5 (muy de acuerdo), afirma literalmente: *"El profesor no explica adecuadamente los contenidos del programa y adolece de una formación evidente"* . Resulta obvio que, en este caso si lo que se quiere es minimizar las diferencias debería de usarse el gráfico B, ya que el A refleja con más precisión lo acontecido, pero sin duda visualiza una mayor diferencia con el resto

de los ítems. Evidentemente, abogamos por el gráfico A en aras al mantenimiento de la veracidad de los datos y conclusiones que se deriva de la investigación.

2.2.2.2. La manipulación del eje de abcisas (o amplitud virtual de X)

Dicho sesgo consiste en añadir categorías de representación en el eje de abcisas que no contienen valor alguno al no existir en realidad. Su efecto más palpable es aumentar considerablemente la pendiente en la estructura que se describe entre las crestas y valles del patrón representado.

GRÁFICO A

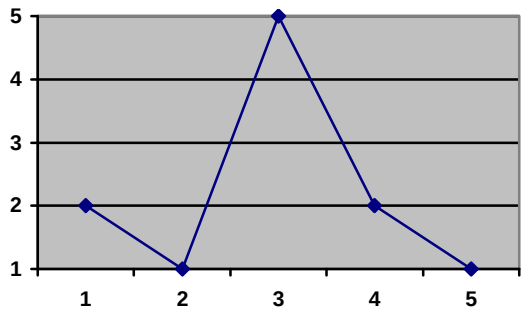
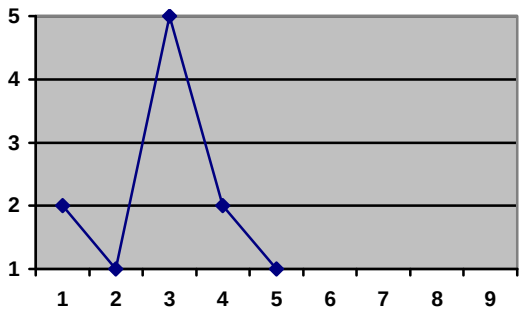


GRÁFICO B



En este caso, la inclusión de cuatro ítems que en realidad no existen y, por tanto, no obtienen puntuación alguna, estrecharía el polígono de frecuencias con el consiguiente aumento de la pendiente del patrón representado.

2.3. Medidas de tendencia central

Existe un conjunto de medidas cuyo cometido es servir como referencia del desempeño conjunto de una colección de valores, es decir, de medida promedio o representativa del resto. Cuando se habla que en España se consume alrededor de 20 litros de cerveza por habitante y año estamos hablando, por supuesto, de un valor promedio que representa al conjunto de los españoles, pero que en cualquier caso, no quiere decir que haya quien no consuma ni una gota mientras otro/as beban, por ejemplo 100 litros. Como valor promedio de un conjunto de sujetos, aspectos… este será más válido, más creíble… cuando se hayan cumplido, al menos, algunos requisitos, como

por ejemplo, la representatividad de los sujetos, aspectos seleccionados de donde se ha extraído la información.

Adoptando el modelo clásico de análisis de datos de tipo descriptivo contemplamos los siguientes estadísticos:

2.3.1. Media aritmética

La media aritmética se define como el sumatorio de valores observados divido por el número de ellos, es decir:

$$\overline{X} = \Sigma xi/N$$

Siguiendo con el ejemplo anterior:

$\overline{X} = 2+1+0+3+2+2+3+1+1...n /20 = 1,7$

2.3.2. Mediana

Por su parte, la mediana se define como la puntuación (xi) que deja detrás y delante el 50% de la distribución de puntuaciones. En este sentido, puede considerarse como el punto medio de una distribución de puntuaciones.

Siguiendo con el ejemplo anterior para su cálculo deberíamos tener en cuenta algunos aspectos:

a) Comprobar si el (N) o número de sujetos o elementos que constituyen la distribución es par o impar.

b) Ordenar la distribución de menor a mayor o viceversa.

Si el número de elementos es par, nuestro caso, tomamos los 2 valores centrales, los sumamos y los dividimos entre 2.

0, 0, 0, 1, 1, 1, 1, 1, 1, *2, 2,* *2, 2, 2, 2, 2, 3, 3, 3, 4*

50% *50%*

Md = 2+2 / 2 = 2

Bien, imaginemos que nos hubiésemos encontrado con una distribución impar de este tipo:

1,1,2,3,3

entonces la Md sería el valor central, o sea, 2

2.3.3. Moda

Se dice que lo que está de moda es aquello que impera en un momento dado. Este invierno se llevará… para la mujer, mientras para el hombre…Tomando el sentido descrito, desde el posicionamiento de la moda textil, podemos afirmar que en estadística la moda es el/los valor/es de la distribución con mayor/es (fi), es decir, el/los que más se repite/n.

Retomando, otra vez, el ejemplo anterior tendríamos que la moda de la distribución sería el valor 2.

Ahora bien, si la distribución fuese esta:

1,1,1,1,2,2,3,3,3,3,4,4,5,6

podemos apreciar que existen dos valores que se repiten por encima de los demás (1 y 3). En ese caso, podemos afirmar que existen dos modas. La distribución sería, pues, bimodal. Puede darse el caso de distribuciones con más de dos modas. Se hablará entonces de una distribución polimodal o multimodal.

2.3.4. Cuestionamiento de la media aritmética como medida representativa del conjunto en algunas ocasiones

En algunas ocasiones en la distribución de puntuaciones existen valores que se alejan bien por defecto, bien por exceso del resto. En estos casos, cuando se presentan valores extremos en la distribución (outliers) y se genera lo que algunos autores denominan: *rough* (desorden) (Tukey, 1977), es aconsejable tomar el valor de la mediana en consideración, ya que es un estadístico más robusto al desorden.

Veamos dos ejemplos diferentes para ilustrar esta situación. Imaginemos que las puntuaciones obtenidas por los niños de dos grupos que han recibido determinados tipos de métodos de lectura son los siguientes:

Grupo A: 1,1,2,3,3
Grupo B: 1,1,2,3,20

En el primer caso (grupo A) se dan las condiciones idóneas para utilizar la media y mediana como medidas de cálculo. En ambos casos, además, la puntuación sería 2. Ese valor promedio representaría adecuadamente al grupo A.

En el segundo caso, hay un valor extremo que puede disparar artificialmente el valor de la media. Así, su valor sería de 8,9, mientras la mediana ascendería 2.

Desde luego 8,9 no representa fielmente a ningún valor de los presentes en la distribución, mientras que el valor 2 (mediana), por lo menos, representa a los valores (1,2 y 3).

2.4. Medidas de dispersión

Si el cometido de las medidas de tendencia central es determinar un valor promedio que represente lo más fielmente al resto, el de las de dispersión o variabilidad es determinar mediante un estadístico cuán homogénea o heterogénea es la distribución de puntuaciones o también a cuánta distancia del centro se encuentran los datos. Evidentemente cuanto más

parecido sean los valores dados más pequeña será la medida de dispersión y vicerversa. Los estadísticos de dispersión fundamentales son:

2.4.1. Amplitud, Rango o Recorrido

La amplitud, rango o recorrido es la diferencia entre valor máximo y el mínimo de la distribución.

Ej: Imaginad las calificaciones obtenidas por 10 niños en un examen de vocabulario: *1,3,4,6,7,5,6,5,8,9*

$A = x_{ima} - x_{imi}$

$A = 9 - 1 = 8$

2.4.2. Desviación media

Por su parte, la desviación media es el cociente entre el sumatorio de cada una de las desviaciones de cada puntuación (xi) respecto de su media en valor absoluto y el valor de N. De ahora en adelante entenderemos por desviaciones la diferencia entre cada valor de la distribución y su media. Si esta diferencia no está elevada al cuadrado, el presente caso, se dice que es una desviación de orden 1. Si está elevada al cuadrado, caso de la desviación típica, se dice que es una desviación de orden 2 y así sucesivamente.

$$DM = \frac{\Sigma \mid xi - \overline{x} \mid}{N}$$

2.4.3. Desviación típica

La desviación típica es el cociente entre la raíz cuadrada del sumatorio de cada una de las desviaciones cuadráticas de cada puntuación (xi) respecto de su media y el valor de N.

$$Sx = \sqrt{\frac{\Sigma (xi - x)^2}{N}}$$

2.4.4. Varianza

La varianza es el cuadrado de la desviación típica. Como podéis apreciar el único cambio en relación a la desviación típica es que ha desaparecido la raíz cuadrada que al pasar al primer término de la ecuación pasa como potencia, es decir, con lo contrario que actuaba en el segundo término de la misma.

$$Sx^2 = \frac{\Sigma\,(xi - x)^2}{N}$$

2.4.5. Coeficiente de variación

Finalmente, el coeficiente de variación es el cociente entre la desviación típica y la media de la distribución. Si se quiere contemplar el resultado en porcentaje basta multiplicarlo por 100.

$$CV = Sx/ \bar{x} * 100$$

Ejemplo: Supongamos que las calificaciones de 5 niños en una prueba de dislexia han sido las siguientes:

1,1,2,3,3

Amplitud = 3 – 1 = 2

DM = Σ (1-2)+ (1-2)+ (2-2)+ (3-2)+ (3-2) = *0,8*

Sx = $\sqrt{\Sigma\,(1-2)^2+ (1-2)^2+ (2-2)^2+ (3-2)^2+ (3-2)^2 / 5}$ = *0,89*

Sx² = Σ (1-2)²+ (1-2)²+ (2-2)²+ (3-2)²+ (3-2)² / 5 = *0,8*

*CV = 0,89 /2 = 0,44 * 100 = 44%*

2.4.6. Cómo interpretar los estadísticos de dispersión

Para la interpretación de las medidas de variabilidad no hay criterios precisos y/o exactos que indiquen formalmente la mayor o menor homogeneidad de la distribución. Con estos precedentes, resulta conveniente tener en cuenta que valores cercanos a 0 implican mayor homogeneidad, mientras valores alejados de este punto todo lo contrario, es decir, mayor heterogeneidad.

2.5. Medidas de posición

El objetivo fundamental de las medidas de posición es incardinar una puntuación referida a un sujeto/objeto en la distribución que conforman ésta y el resto de puntuaciones, es decir, establecer qué porcentaje y cuántos sujetos/objetos se sitúan por debajo y por encima de la misma.

Existen diferentes tipos de medidas de posición, genéricamente denominadas cuantiles. Los cuantiles más usados son los percentiles o centiles, los deciles y los cuartiles. Cada uno de ellos se caracteriza por unos rasgos determinados, que en la tabla siguiente trataremos de sintetizar.

Cuantil	Definición	Número de partes en que se constituye
Centil	Medida de posición que divide la distribución en 100 partes cada una conteniendo una centésima parte de las observaciones	Del C_1.... al C_{99}
Decil	Medida de posición que divide la distribución en 10 partes cada una conteniendo al 10% de las observaciones	Del D_1.... al D_9
Cuartil	Medida de posición que divide la distribución en 4 partes cada una de ellas conteniendo un cuarto (25%) de las observaciones	Del Q_1.... al Q_3

Así por ejemplo, podemos afirmar que tras el Q_3 se encuentran el 75% de las observaciones, o que delante del mismo están el 25%.

Evidentemente, se dan una serie de equivalencias entre los diferentes cuantiles que también hemos contemplado en la siguiente tabla:

	D_1	C_{10}
	D_2	C_{20}
Q_1	→	C_{25}
	D_3	C_{30}
	D_4	C_{40}
Q_2	D_5	C_{50}
	D_6	C_{60}
	D_7	C_{70}
Q3	→	C_{75}
	D_8	C_{80}
	D_9	C_{90}

Las fórmulas habituales para el cálculo de cuantiles que pueden encontrarse en cualquier manual de estadística aplicada a las ciencias sociales son para casos en que la distribución está organizada en intervalos. Como en nuestros procedimientos de cálculo no contemplamos esa posibilidad utilizaremos las siguientes expresiones para su cálculo:

$Pk = (n+1) * p$

siendo:

n = número total de observaciones

p= proporción del cuantil

En caso de que la posición calculada no fuese exacta, es decir, se obtengan decimales deberá utilizarse esta expresión de interpolación:

$Pk = (1-\alpha) * xi1 + (\alpha) * xi2$

siendo:

α: cuantía decimal o parte de fracción de la posición determinada

xi1: valor de la primera observación que contiene la posición del percentil en cuestión o posición más cercana por defecto

xi2: valor de la segunda observación que contiene la posición del percentil en cuestión o posición más cercana por exceso

Ejemplo: El número de faltas cometidas en un dictado por nueve niños con trastornos de disgrafía han sido las siguientes:

13,13,14,15,15,15,16,17,17

A partir de estos datos calcular el valor del percentil 25

Para calcular dicho percentil debemos desplegar los siguientes pasos:

1º Se organizan las observaciones de menor a mayor indicando sus frecuencias absoluta y acumulada, o sea:

Xi	*Fi*	*Fa*
13	2	2
14	1	3
15	3	6
16	1	7
17	2	9

2º Se calcula la posición del P25 en la distribución de observaciones mediante la siguiente expresión:

Lugar del P25 = (n+1) * p

Aplicando esta expresión P25 = (9+1) * 0,25 tendríamos que el lugar que ocupa el P25 en la distribución de nueve observaciones es 2,5º. Al no ser un lugar exacto se deberá interpolar mediante la ecuación:

P25 = (1-α) * *xi1* + (α) * *xi2*

Aunque antes de operar mediante esta expresión situamos el P25 en la distribución de observaciones sirviéndonos para ello de la tabla de frecuencias:

Xi	*Fi*	*Fa*
13	2	2
14	1	3
15	3	6
16	1	7
17	2	9

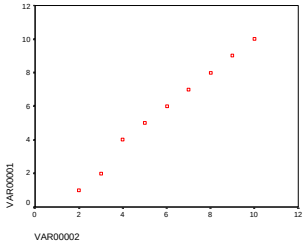
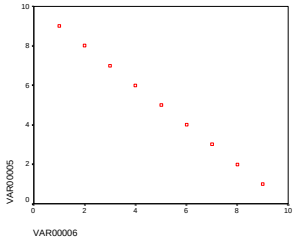
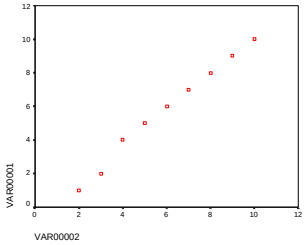
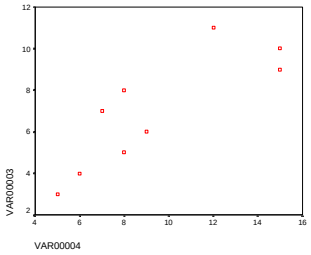
Posición del P25

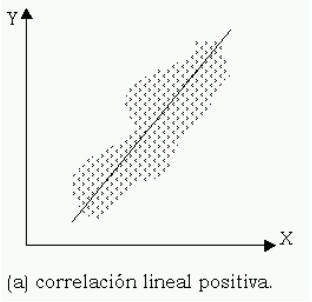
Podemos apreciar como la posición 2,5º está contenida en las frecuencias acumuladas 2 y 3 cuyas observaciones de referencia son 13 y 14. Si aplicamos la fórmula de interpolación tendremos que:

P25 = (1-0,5) * *13* + (0,5) * *14* = *13,5*

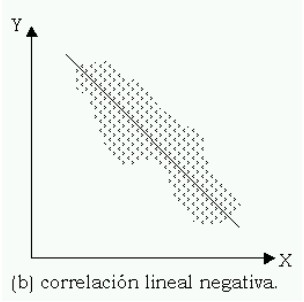
2.6. La correlación

La relación entre dos variables nos conduce a un nuevo concepto: correlación. Ésta puede ser representada en diagramas de dispersión e informa acerca de la *forma*, *dirección* e *intensidad* de la relación entre dos variables, en ningún caso sobre los efectos de una sobre otra (ausencia de causalidad, pero cierto grado de predicción). Con Yela (1994:247) podemos afirmar que en referencia a la *forma* ésta puede ser una línea recta si la relación es lineal, es decir, si las diferencias entre los valores de una variable son proporcionales directa (+) o inversamente (-) a las diferencias entre los correspondientes a la otra; o una *curva*, cuando la relación, no siendo lineal, es, o bien monotónica (incrementos iguales en una variable corresponden a incrementos crecientes, o decrecientes en la otra), o bien no monotónica, cuando hay cambios de dirección en la curva. Con relación a la *dirección* puede ser variable o constante, bien (positiva), cuando a incrementos o decrementos de la variable “A” corresponden incrementos o decrementos de la variable “B”, o negativa cuando a incrementos de la variable “A” corresponden decrementos en la variable “B” y viceversa. Con respecto a la *intensidad* de la relación se manifiesta en la dispersión de los datos en torno a la línea (recta de regresión) y suele expresarse mediante valores de coeficientes de correlación que oscilan entre –1 y 1 pasando por 0. Valores cercanos a 0 denotan ausencia de correlación y, por ende, independencia, valores cercanos a –1 ó 1 indican una correlación de gran intensidad y, por tanto, una fuerte relación.

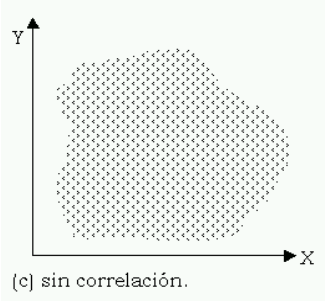
DIRECCIÓN Y/O SIGNO	
Positiva	Negativa
	
	
Mayor intensidad	Menor intensidad
INTENSIDAD	



(a) correlación lineal positiva.



(b) correlación lineal negativa.



(c) sin correlación.

2.6.1. El coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es el coeficiente de su modalidad más usado. Sin embargo, su utilización está sujeta a una condición imprescindible: las dos variables tienes que estar medidas en intervalo. Existen varias ecuaciones para determinar el valor del coeficiente de Pearson. Te proponemos dos de las más utilizadas; la clásica en formato largo y la abreviada basada en la covarianza y desviaciones típicas de las variables *x* e *y*. Téngase en cuenta que la covarianza se basa en el sumatorio del producto cruzado de las desviaciones partido el número de sujetos/elementos objeto de análisis, es decir:

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Por su parte, las dos fórmulas que vamos contemplar son las siguientes:

$$r_{xy} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)}\sqrt{\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

$$= \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{n(\sum X^2) - (\sum X)^2}\sqrt{n(\sum Y^2) - (\sum Y)^2}}$$

$$r = \frac{S_{xy}^2}{S_x S_y}$$

La primera fórmula es la denominada fórmula larga, aparatosa a primera vista, pero sencilla cuando se tabulan y organizan los datos desde unas determinadas coordenadas que a continuación explicitaremos. La segunda fórmula se basa en la covarianza divida entre el producto de las desviaciones típicas de las variables “x” e “y”.

La característica fundamental de este índice es que mide la existencia de una relación lineal entre dos variables medidas en escala de intervalo. El valor de este coeficiente de correlación varía en el intervalo [-1, +1], de tal forma que un coeficiente de correlación de Pearson igual o cercano a 0 indica una independencia total o amplia entre las dos variables y así puede decirse que cuando una de ellas varía esto no influye en absoluto, o acaso levemente, en el valor que pueda tomar la segunda variable. Por su parte, un valor de correlación igual o cercano a (–1) indica una dependencia total o fuerte entre las dos variables, denominada *relación inversa*, de manera que cuando una de ellas aumenta la otra disminuye y viceversa. Finalmente, un coeficiente de correlación igual o cercano a (+1) indica una dependencia también total o fuerte entre las dos variables, denominada *relación directa*, de manera que cuando una de ellas aumenta la otra también aumenta y cuando disminuye también lo hace la otra.

Imaginemos que un licenciado en Psicopedagogía desea averiguar que relación existe entre el número de horas de estudio el fin de semana previo a un examen y las calificaciones obtenidas en el mismo. Para ello examina el caso de diez alumnos que dicen haber estudiado las siguientes horas habiendo obtenido también las siguientes calificaciones

Horas de estudio (X)	*Calificaciones (Y)*
2	*3*
2	*3*
3	*3*
4	*4*
5	*5*
6	*7*
7	*7*
7	*7*
9	*8*
10	*9*
Σ 48	*Σ 56*

La relación que pueden guardar ambas variables puede determinarse mediante el coeficiente de correlación de Pearson. Para ello es condición indispensable que ambas variables estén medidas en escala de intervalo (se cumple dicho supuesto). Por tanto, sólo queda aplicar la dichosa y farragosa ecuación antes propuesta. Para ahorrarte sufrimiento te proponemos que organices los datos de la siguiente forma:

Horas de estudio (X)	*Calificaciones (Y)*	*X²*	*Y²*	*X*Y*
2	*3*	*4*	*9*	*6*
2	*3*	*4*	*9*	*6*
3	*3*	*9*	*9*	*9*
4	*4*	*16*	*16*	*16*
5	*5*	*25*	*25*	*25*
6	*7*	*36*	*49*	*42*
7	*7*	*49*	*49*	*49*
7	*7*	*49*	*49*	*49*
9	*8*	*81*	*64*	*56*
10	*9*	*100*	*81*	*90*
Σ 48	*Σ 56*	*Σ 373*	*Σ 360*	*Σ 348*

Ahora puedes aplicar la fórmula:

$$r_{xy} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)}\sqrt{\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

$$= \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{n(\sum X^2) - (\sum X)^2}\sqrt{n(\sum Y^2) - (\sum Y)^2}}$$

Sustituyendo cada miembro por su valor tendríamos:

$$Rxy = 9 * (348) - (48*56) / \sqrt{9*(373)- (48)^2} * \sqrt{9*(360)-(56)^2} = 0.97$$

Interpretación del coeficiente obtenido:

1. En primer lugar la correlación obtenida es positiva lo cual indica que mayor numero de horas de estudio mayor calificación en el examen. Era de prever, no obstante, que se obtuviese tal resultado.
2. El valor obtenido está muy cercano a 1. Ello significa que la relación entre ambas variables es muy importante y que, por tanto, no sería descabellado implementar un análisis de regresión con el objetivo de predecir valores de la variable criterio, en este caso calificación en el examen. Pero tranquilo/a estimado/a alumno/a eso no es objeto de la siguiente obra y no se va a dar.

2.6.2. El coeficiente de correlación de Rho de Spearman

La correlación de Spearman, por su parte, es una aplicación del coeficiente de correlación de Pearson a “n” pares de observaciones cuyos valores son números de orden.

Para su cálculo se procede según los pasos siguientes:

- Se transforman los valores originales por sus rangos. Esta transformación se realiza independientemente para cada variable.
- Se obtiene, para cada sujeto, la diferencia, en valor absoluto, entre los rangos.

Estas diferencias permiten obtener el valor del coeficiente de correlación de Spearman, por medio de la siguiente expresión:

$$r_s = 1 - \frac{6(\sum d^2)}{n(n^2 - 1)}$$

La interpretación de "r_s" es idéntica a la del coeficiente de correlación de Pearson. Su valor oscila en el intervalo [-1,+1] siendo el valor 0 indicador de independencia entre las dos variables.

El valor +1 aparece cuando los rangos asignados a los dos valores de un mismo sujeto son iguales, y el valor -1 aparece cuando los rangos asignados son opuestos.

x	1º 3º 4 º2º 6º 5º
y	1º 3º 4º 2º 6º 5º

r_s= +1

x	1º 3º 4º 2º 6º 5º
y	6º 4º 3º 5º 1º 2º

r_s= -1

La expresión anterior sólo proporciona el resultado correcto cuando todas las observaciones sean diferentes y, por tanto, le correspondan rangos diversos a cada una de ellas.

Cuando haya observaciones empatadas, el coeficiente de Spearman se debe obtener por medio de la siguiente relación:

$$r_s = \frac{A + B - D^2}{2\sqrt{AB}}$$

donde los valores de A y B se obtienen a través de:

$$A = \frac{N^3 - N - T_1}{12}$$

$$B = \frac{N^3 - N - T_2}{12}$$

donde los valores T_1 y T_2 corresponden al grupo 1 y grupo 2 respectivamente, obtenidos por medio de $T_i = t^3 - t$, siendo **t** el número de observaciones ligadas (empates) en el rango i:

$$T_j = \sum(t_i^3 - t_i)$$

El psicopedagogo de un centro escolar ha pasado dos test que miden dos variables diferentes. Los resultados obtenidos están medidos en escala de intervalo y son los siguientes:

X	*y*
10	*13*
12	*17*
16	*15*
14	*15*
10	*16*
12	*12*

Sin embargo, para tratar de determinar si entre ambas variables existe relación decide implementar el coeficiente Rho de Spearman, para lo cual tendrá que transformar las puntuaciones medidas en escala de intervalo en rangos de orden.

Así pues, en primer lugar, se realiza la transformación de valores originales en números de orden o rangos, por separado:

Valores	10	10	12	12	14	16	12	13	15	15	16	17
Rango	1º	2º	3º	4º	5º	6º	1º	2º	3º	4º	5º	6º
Rango asignado	1,5º		3,5º		5º	6º	1º	2º	3,5º		5º	6º

A partir de los números de orden se genera la variable D= x-y, cuyos valores, así como sus cuadrados, aparecen en la tabla siguiente:

x	y	D	D^2
1.5º	2º	0.5	0.25
3.5º	6º	2.5	6.25
6º	3.5º	2.5	6.25
5º	3.5º	1.5	2.25
1.5º	5º	3.5	12.25
3.5º	1º	2.5	6.25
			$\sum$ 33.5

Ya que hay empates en los rangos procederemos a realizar la corrección:

$$T_1 = (2^3 - 2) + (2^3 - 2) = 12$$

$$T_2 = (2^3 - 2) = 6$$

Los valores de A y B se obtienen a través de:

$A = N^3 - N - T_1 / 12 = 6^3 - 6 - 12 / 12 = 16.5$

$B = N^3 - N - T_2 / 12 = 6^3 - 6 - 6 / 12 = 17$

Así pues, el valor de la correlación de Spearman es:

$$r_s = \frac{A + B - D^2}{2\sqrt{AB}} = \frac{16.5 + 17 - 33.5}{2\sqrt{16.5 \times 17}} = 0$$

El resultado obtenido implica que la relación existente entre las variables “x” e “y” es totalmente nula. En ese sentido, podemos afirmar que son dos variables sin relación alguna o totalmente independientes.

2.6.3. Coeficientes de correlación basados en el chi cuadrado

En ocasiones las dos variables que se están correlacionando poseen una naturaleza claramente nominal. Para esos casos se contemplan numerosos coeficientes de correlación basados en una prueba de contraste de hipótesis denominada chi cuadrado. Por su importancia y uso habitual destacamos los siguientes:

a) El coeficiente phi o cuádruple

Como los otros dos que vamos a contemplar, este coeficiente se utiliza cuando las dos variables correlacionadas son de naturaleza nominal. La primera ecuación que proponemos sólo es válida para el caso de cruces 2x2, es decir, que ambas variables tengan un máximo de dos niveles cada una. Dicha ecuación es la siguiente:

$$\phi = (A*D) - (B*C) / \sqrt{(A+B)*(A+C)*(C+D)*(B+D)}$$

donde

A, B, C y D son frecuencias observadas correspondientes a las celdillas pertenecientes a las intersecciones de los diferentes niveles, o sea:

	J		
I		J_1	J_2
	I_1	A	B
	I_2	C	D

Otra fórmula, ésta sí compatible para cruces superiores a 2x2, es la siguiente:

$$\phi = \sqrt{\chi^2/N}$$

donde

χ^2 = valor de la prueba de contraste de hipótesis con el mismo nombre

N = número de sujetos objeto de análisis

b) Coeficientes de contigencia y "V" de Cramer

Otros coeficientes son el coeficiente de contingencia, así como el "V" de Cramer. Mientras el primero (coeficiente de contingencia) tiene en cuenta igualmente el valor de chi cuadrado y cuya expresión es la siguiente:

$$C = \sqrt{\chi^2 / \chi^2+N}$$

el segundo (V de Cramer) se basa en el valor de phi divido entre los grados de libertad mínimos de fila y columna. La fórmula de cálculo es la siguiente:

$$V = \sqrt{\phi^2 / \text{mín } (I-1),(J-1)}$$

2.6.4. La regresión estadística

Un caso especial de correlación resulta ser la predicción de una variable, llamémosle criterio o dependiente (*y´*) a partir de otra variable predictora o independiente (*x*) tomando como fundamento la correlación (rxy) que guardan ambas variables. Nos encontramos ante la regresión lineal simple, cuya ecuación matemática se define como:

$$Y' = \alpha + \beta * xi + \varepsilon$$

o también:

$$Y' = a + b * xi + e$$

donde:

Y´: valor criterio

a: intercepto o punto de corte de la recta de regresión con el eje de ordenadas *y*

b: pendiente o tangente de la recta de regresión

xi: variable predictora

e: desviación o inexactitud del ajuste que a su vez se define.

2.6.4.1. Significado y ecuaciones de cálculo de las constantes "a" y "b"

Como hemos explicitado anteriormente el coeficiente "b", también llamado tangente o pendiente de la recta de regresión, indica los incrementos de la variables dependiente (y) cuando la variable independiente (x) aumenta en una unidad. Servirá como un indicador del sentido de asociación entre ambas variables, de tal forma que un b>0 nos indicará una relación directa entre ellas (a mayor valor de la variable explicativa, el valor de la variable dependiente *y* aumentará), b<0 delatará una relación de tipo inverso, mientras que b=0 nos indica que no existe una relación lineal clara entre ambas variables. Una fórmula, entre las que se contemplan para su cálculo, puede ser la siguiente:

$$\hat{b} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Apréciese que el numerador de la expresión está formado por el sumatorio de los productos cruzados de las variables "x" e "y" y como denominador contemplamos el momento de orden dos de la variable "x". En el primer caso, se trata de calcular las distancias desde cada puntuación de las variables "x" e "y" a sus respectivas medias, mientras en el segundo también determinar las distancias desde cada punto de la distribución de puntuaciones "x" a su media y elevarlo al cuadrado.

En cuanto a la constante "a" o intercepto hace referencia al lugar por donde la ecuación de regresión corta con el eje de ordenadas "y". En este sentido, podemos afirmar que el coeficiente "a" indica el valor de "y" cuando la variable "x" toma el valor 0. Representa, por tanto, la influencia

de otras variables que no hemos tenido en cuenta al analizar la variable. Una vez calculado “b” será fácil el cálculo de “a” a partir de la siguiente expresión:

$$\hat{a} = \overline{Y} - \hat{b}\overline{X}$$

BIBLIOGRAFÍA:

Darrell, H. y Geis, I. (1954/1993). *How to lie with statistics*. New York. WW. Norton & Company/Paperback.

Etxeberria Murgiondo, J. y Tejedor Tejedor, F. J. (2005) *Análisis descriptivo de datos en educación*. Madrid: La Muralla.

Gil Flores, J. Rodríguez Gómez, G. y García Jiménez, E. (1995). *Estadística básica aplicada a las Ciencias de la Educación*. Sevilla: Kronos.

Gil Flores, J. Rodríguez Gómez, G. y García Jiménez, E. (1996). *Problemas de estadística básica aplicada a las Ciencias de la Educación*. Sevilla: Kronos.

Monmonier, M. (2001). *Bushmanders and Bullwinkles: How Politicians Manipulate Electronic Maps and Census Data to Win Elections*. Chicago. University of Chicago Press.

Salvador Figueras, M y Gargallo, P. (2003): "Análisis Exploratorio de Datos", [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [28 de junio de 2006]

Tufte, H. (2003) (2ª edición). *The visual display of quantitative information*. Cheshire. Graphics Press.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading. Addison Wesley.

Yela, M. (1994). Análisis de datos, en García Hoz, V. (Dir.). *Problemas y métodos de investigación en educación personalizada*. Madrid. Rialp, pp. 223-254.