Apuntes de Estadística para profesores

Curso 2006/2007

Concepción Bueno García
Tomás Escudero Escorza

Instituto de Ciencias de la Educación
Universidad de Zaragoza

Capítulo 1. Conceptos generales

1.- Introducción

Las dos grandes funciones de la Estadística (descripción de datos y realización de inferencias) reflejan la propia historia del desarrollo de esta ciencia. La Estadística actual es el producto del encuentro y la propia fecundación de dos ramas distintas del saber, la antigua estadística y el cálculo de probabilidades, que se encontraron en el siglo XIX. Etimológicamente, la palabra estadística procede de la palabra estado. Ya en la antigüedad los romanos y los egipcios hicieron intentos por tener un conocimiento preciso del número de sus habitantes y de sus posesiones, es decir, por conocer el estado de sus naciones (de ahí la raíz del término). Para ello hicieron recolecciones de datos que posteriormente tenían que resumir de una forma comprensiva y que permitiera proporcionar informaciones útiles. Este tipo de estudios dio lugar a la estadística descriptiva cuya misión consiste en describir situaciones y procesos dados; para ello se sirve de tablas, representaciones gráficas, proporciones, números índice y medidas típicas.

Sin embargo las conclusiones extraídas se agotan en el propio conjunto de datos observados, pues el objetivo consistía en hacerse una idea clara de lo que había, y lo que había se contaba y se medía. Lo que posibilitó el cálculo de probabilidades fue, precisamente, el desarrollo de un conjunto de métodos para extrapolar las conclusiones a entidades no observadas. Es decir, proporcionó el instrumento adecuado para poder hacer inferencias acerca de grandes cantidades de observaciones potenciales a partir de unas pocas observaciones reales. Estas técnicas tuvieron su fundamento en el desarrollo de la curva normal por Gauss, en su aplicación por Galton a los problemas de herencia,etc. Sin embargo los auténticos fundadores de estas técnicas fueron Karl Pearson (1857-1936) y Sir Ronald Fisher (1890-1962). Así se ha desarrollado la estadística analítica o inferencial basada en la teoría de probabilidades que trata de obtener leyes generales a partir de la observación de algunos datos. Precisamente este fundamento probabilístico condiciona el que los resultados obtenidos se vean sujetos a unos márgenes de error.

Ahora se puede dar una definición de Estadística en la que aparecen algunos términos no definidos lo cual no impedirá entender su significado.

Estadística es la ciencia que se ocupa de la ordenación y análisis de datos procedentes de muestras, y de la realización de inferencias acerca de las poblaciones de las que éstas proceden.

2.- La Estadística como herramienta para el profesor

Dentro del ambiente educativo la Estadística es necesaria al menos para llevar a cabo estas cuatro tareas:

1ª Lectura de literatura profesional

La investigación en Ciencias de la Educación emplea la Estadística como herramienta habitual en la realización de cualquier experimento. Por tanto, el profesor que

quiera estar al día respecto a la enseñanza de su disciplina debe estar en condiciones de poder comprender textos de investigación en Ciencias de la Educación.

2ª Conocimiento de la clase

El profesor se enfrenta a la tarea de la educación de unos alumnos ubicados en una clase, centro escolar y contexto social concreto, que van a interaccionar con sus características personales. El conocimiento profundo de este contexto en el que está involucrado el alumno resulta de vital importancia para el educador y no será posible sin el análisis estadístico de los datos individuales de los elementos del contexto.

3ª Diagnosis didáctica

El profesor, a la hora de tomar decisiones acerca de sus alumnos, contará con el apoyo del análisis comparativo de la situación relativa de cada individuo en su clase, distintas asignaturas y distintas variables psico-sociológicas. También la propia actividad del profesor puede verse mejorada tras un análisis del rendimiento escolar del grupo en su conjunto. Estas tareas requieren tratamientos estadísticos simples de los datos de los alumnos.

4ª Investigación y predicción

El profesor puede estar interesado en averiguar si una nueva técnica didáctica es realmente más efectiva de cara al rendimiento de sus alumnos que la usada por él hasta ahora. O en saber el efecto que producen variables familiares, rasgos psicológicos en la destreza del alumno en realizar tal o cual tarea... Este tipo de trabajos requieren el uso de métodos estadísticos.

3.- Primeros conceptos

Normalmente, el investigador desea conocer un parámetro o característica de los elementos de una población. Esta población suele ser demasiado extensa (razones económicas) o poco definida (votantes) como para estudiarla al completo. Entonces se lleva a cabo una selección (o muestreo) del que se obtiene una muestra de elementos que sea una representación de la misma. Se mide a estos sujetos de la muestra la característica buscada y se calcula el valor de esa característica para esa muestra. Esa medida es un estadístico muestral que es, a su vez, una estimación del parámetro de la población.

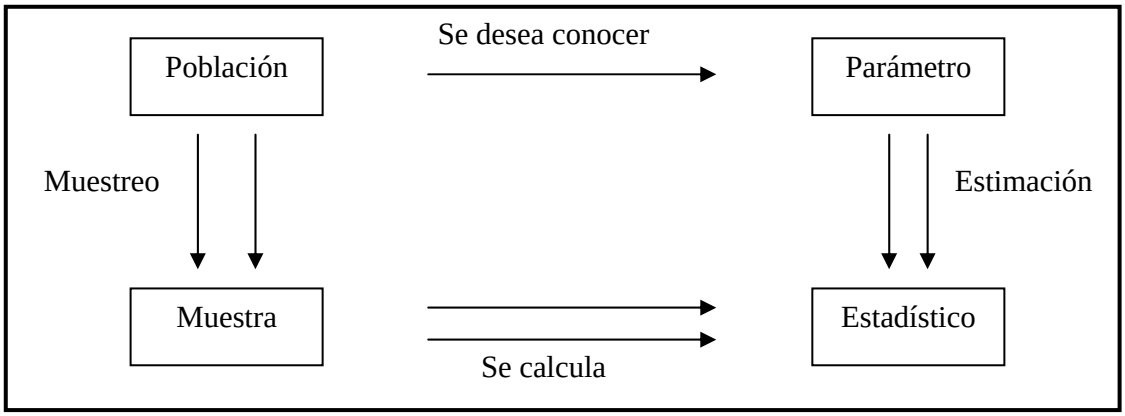

Figura 1: Esquema de la relación entre los conceptos de población, muestra, parámetro y estadístico

En la Figura 1 aparece la relación existente entre estos conceptos. La flecha doble indica el camino seguido habitualmente por el investigador.

La situación descrita anteriormente nos lleva a dar las siguientes definiciones:

- Población estadística es el conjunto de todos los elementos que cumplen una o varias características o propiedades.
- Una muestra es un subconjunto de los elementos de una población.
- Un parámetro es una propiedad descriptiva de la población.
- Un estadístico es una propiedad descriptiva de la muestra.

Pongamos ahora algunos ejemplos que nos permitan identificar los conceptos que hemos definido:

1º.-Supongamos que estamos interesados en saber cómo es la actitud de los estudiantes de la Universidad de Zaragoza hacia la práctica deportiva continuada. La población es en estos momentos de alrededor de 50.000 estudiantes. Parece obvio que preguntarles a todos ellos acerca de esta actitud resultaría bastante costoso por lo que decidimos seleccionar algunos de ellos para realizar la consulta. Con los resultados obtenidos a partir de estos alumnos seleccionados (muestra), podremos generalizar sobre la actitud de todos los alumnos de la UZ. En este estudio uno de los parámetros podría ser el tanto por ciento de alumnos de la UZ que tienen una actitud positiva hacia la práctica deportiva continuada. Este parámetro sería estimado por el valor del correspondiente estadístico en la muestra que hemos seleccionado, es decir, por el tanto por ciento de alumnos con actitud positiva en la muestra.

2º.-Un profesor de historia quiere medir la eficacia del uso sistemático de la hemeroteca para la enseñanza de la Historia Contemporánea de 3º de ESO. Este profesor no puede realizar un estudio que implique a todos los alumnos de 3º de ESO de España, así que decide hacer un experimento con dos clases de las que él mismo es profesor. En primer lugar confecciona una prueba de conocimientos previos de un tema de la asignatura y la pasa a sus alumnos. En la clase A, aparte de la habitual lección magistral, pide a sus alumnos que lean algunos periódicos que reflejen algún aspecto del tema. En la clase B se limita a dar su lección magistral. Cuando acaba la explicación del tema examina a sus alumnos. Ahora, con ayuda de la Estadística, puede responder a algunas preguntas como:

¿Obtienen mejores notas los alumnos de la clase A que los de la clase B?
¿Partían del mismo nivel de conocimientos previos?
¿Podría este hecho influir sobre los resultados del experimento?
¿Se pueden generalizar los resultados del experimento de este profesor?

3º.-Queremos estudiar el procesamiento de información en tareas simples, para lo cual se ha utilizado tradicionalmente como medida el tiempo de reacción. En este caso estamos interesados en el tiempo que necesita un sujeto en concreto para realizar una tarea. Para ello, se presentan al sujeto en cada ensayo uno de dos posibles estímulos. Ante uno de ellos el sujeto deberá presionar un botón tan rápido como pueda, mientras que no debe dar respuesta alguna ante el otro. Para hacernos una idea global de cómo realiza la tarea nuestro sujeto, decidimos administrarle 30 ensayos con lo que obtenemos 30 datos, cada uno de ellos representando el tiempo invertido por el sujeto en cada uno de los ensayos. En este caso la población la componen todas las

realizaciones de la tarea que potencialmente puede realizar nuestro sujeto experimental. La muestra la constituyen los treinta ensayos. El estadístico podría ser la media del tiempo empleado en los ensayos.

4º.-Tratamos de hacer un sondeo acerca del resultado de un referéndum que se celebrará próximamente en España. La población en este caso está poco definida porque no todas las personas que pueden votar lo hacen realmente y, además, resultaría muy costoso preguntar a todos los votantes por su intención de voto por lo que seleccionamos 3000 españoles (muestra) que consideramos representativos y les preguntamos por el sentido de su voto ante el referéndum. El porcentaje de individuos de la población que responderían SI es un parámetro. El porcentaje de nuestra muestra que responde SI es un estadístico, una estimación de ese parámetro.

4.- Variables y su clasificación

A través de estos ejemplos se ve que cuando estudiamos las entidades que conforman una población nos interesamos por algunas de las propiedades de sus elementos, y esas propiedades adoptan distintas variedades:

- Una característica es una propiedad o cualidad de un individuo.
- Una modalidad es cada una de las maneras como se presenta una característica.

Podemos señalar como ejemplo algunas de las características y modalidades de los alumnos de Enseñanza Secundaria:

- Rendimiento académico en las asignaturas cursadas que adopta distintas modalidades, normalmente son valores entre 0 y 10.
- Sexo que adopta dos modalidades: varón y mujer
- Lugar de procedencia
- Motivación ante la asignatura
- Empleo del tiempo de ocio

Ahora estamos en condiciones de definir el concepto de variable como la característica de los elementos de una población que toma ó puede tomar valores distintos en cada uno de ellos. En contraposición a éste aparece el concepto de constante que es una característica de la población que sólo puede tomar un valor para todos los elementos de la población.

Las variables que aparecen en los ejemplos son:

- Actitud de los estudiantes de la UZ hacia la práctica continuada del deporte.
- Nota obtenida por los alumnos en la prueba de conocimientos previos.
- Nota obtenida por los alumnos en la prueba final.
- Tiempo de realización de la tarea.
- Voto del referéndum.

Las constantes que aparecen en los ejemplos son:

- En el primero el nivel de estudios de los encuestados (todos son universitarios).
- En el segundo los estudios cursados (3º de ESO).

- En el cuarto la nacionalidad de los encuestados (española).

Las variables se pueden clasificar según el número de valores que puedan tomar como variables discretas y variables continuas.

Una variable continua es la que puede tomar todos los valores de un intervalo. Por ejemplo: el peso, la talla, el tiempo empleado en la ejecución de una tarea, la duración de un suceso, etc.

Cuando tratamos con variables continuas y las definimos como variables que pueden tomar cualquier valor, tenemos que tener en cuenta la precisión del instrumento de medida que estamos usando. En realidad una variable continua nunca puede medirse con total precisión, no podemos conocer su valor exacto, sino su valor informado, que es el que nos proporciona el instrumento de medida. Por ejemplo, si utilizamos el metro como unidad de medida, con aproximación de centímetros, cuando decimos que una persona mide 1,65 queremos decir que mide entre 1,645 y 1,655 .

Una variable discreta es aquella que adopta valores aislados. Ejemplo: número de asignaturas aprobadas en el curso pasado, número de alumnos de una clase, sexo, sentido del voto en unas elecciones, nivel socioeconómico, etc.

5.- Medición y escalas

Llamamos medición al proceso de atribuir números a las variables. El conjunto de reglas o modelos desarrollados para la asignación de números a las variables es lo que se denomina escala. La clasificación de las escalas más usada es la propuesta por Stevens (1946) que divide las escalas en: nominales, ordinales, de intervalo y de razón.

Escala nominal: nos permite identificar sujetos como "iguales" o "diferentes". Usando una escala nominal podemos decidir si un sujeto es igual o diferente a otro, pero no podemos establecer relaciones de orden respecto a esa característica, ni relaciones de cantidad ni de diferencia. Por ejemplo: si medimos el color de los ojos podemos establecer la siguiente escala: A $\rightarrow$ azul, V $\rightarrow$ verde, M $\rightarrow$ marrón y N $\rightarrow$ negro. No podemos ordenar los sujetos de mayor a menor o viceversa, simplemente podemos asegurar si dos sujetos tienen el mismo o distinto color de ojos. Otros ejemplos: nacionalidad, sexo, profesión. A este tipo de variables medidas con escala nominal se les puede asignar a cada categoría cualquier tipo de símbolos. En el ejemplo hemos asignado letras pero podíamos haber optado por números: 1 $\rightarrow$ azul, 2 $\rightarrow$ verde, 3 $\rightarrow$ marrón y 4 $\rightarrow$ negro.

Escala ordinal: Esta escala no sólo permite la identificación y diferenciación de los sujetos sino que además permite establecer relaciones del tipo "mayor que" o "menor que". Es decir, de los sujetos se puede decir cual presenta una mayor o menor magnitud de la característica medida, los objetos se pueden ordenar. Ejemplo: nivel de estudios se puede asignar 1 a estudios primarios, 2 a estudios secundarios, 3 a estudios universitarios. Podemos ordenar a los sujetos según el nivel de estudios, el valor 3 es mayor que el 2 y el 1. Aunque no podemos afirmar que la diferencia existente entre el 2 y el 1 sea la misma que la que existe entre el 3 y el 2. Ni que el que tenga nivel 3 tenga 3 veces más de nivel de estudios que el que tiene nivel 1. Otros ejemplos de escala ordinal: posición relativa en la clase, escala de dureza de los minerales.

Escala de intervalo: Con esta escala, además de poder identificar un objeto y establecer relaciones del tipo mayor que y menor que, también podemos hacer afirmaciones acerca de las diferencias en la cantidad del atributo de unos y otros objetos. Es decir, disponemos de una unidad de medida, aunque en este caso el cero sea un punto arbitrario en la escala. Es decir, no indica ausencia total de la cantidad de atributo. Un ejemplo típico es el calendario, podemos afirmar que ha transcurrido el mismo tiempo entre 1960 y 1966 que entre 1980 y 1986 porque contamos con una unidad de medida llamada año. Pero no podemos afirmar que hasta el año 1000 haya pasado el doble de tiempo que hasta el año 500, porque el valor cero no representa el comienzo del tiempo sino que, en nuestro calendario se eligió el año del nacimiento de Cristo como año 1. Otros ejemplos: la medición de las temperaturas en grados centígrados, la escala de los test de inteligencia, las calificaciones escolares.

Escala de razón: También se llama de proporción o de cociente. Además de las características de las otras tres escalas, contamos con una unidad de medida con cero absoluto, es decir, que significa ausencia del atributo o característica medida. Por ejemplo, la longitud, podemos afirmar que un objeto que mide 10 cm. tiene el doble de longitud que uno que mide 5 cm. Otros ejemplos: peso, duración de un suceso, temperatura en grados Kelvin (que sí tiene cero absoluto).

Una vez descritas estas escalas podemos volver a clasificar las variables según la escala usada para medirlas, es decir, podemos hablar de variables nominales, ordinales, de intervalo y de razón.

También se pueden clasificar atendiendo al tipo de información que proveen en cualitativas y cuantitativas.

Variables cualitativas son aquellas que se miden según una escala nominal u ordinal. Informan más bien de una cualidad del sujeto: sexo, color de ojosnivel socioeconómico, nivel cultural, dureza de los minerales.

Variables cuantitativas son aquellas que se miden según una escala de intervalo o de razón. De alguna forma dan cuenta de la cantidad de atributo o característica que el individuo posee. Por ejemplo: peso, talla, temperaturas, número de asignaturas aprobadas, calificación en la última evaluación de la asignatura X.

En la mayoría de las investigaciones educativas las variables manejadas se miden con escalas nominales, ordinales o de intervalo. Son escasamente utilizadas las de razón o proporción, salvo que se use, por ejemplo, el tiempo utilizado para ejecutar una tarea.

Capítulo 2. Organización y representación de datos.

1.- Distribución de frecuencias, histograma y polígono de frecuencias.

En el capítulo anterior ha quedado claro el hecho de que la Estadística trabaja con datos de muy diversa índole:

- Datos que provienen de la medición de variables: peso, talla, tiempo empleado en realizar una tarea, rendimiento académico, etc.

- Datos que son frecuencias de categorías, que provienen de un proceso de conteo: número de nacimientos, número de matriculados en un curso, número de escolarizados, etc.

- Datos que reflejan porcentajes, probabilidades: porcentaje de aprobados en un centro, probabilidad de pertenecer a un grupo dentro de un determinado curso

- Ratios o números índice, que son números que provienen de un cociente: el índice de precios al consumo, el IQ = (edad mental / edad cronológica) · 100, ratio alumnos-profesor, tasa de aprobados por especialidad.

En general, una vez que el profesor o investigador ha recabado información acerca de sus alumnos ó de la muestra elegida en su caso, a través de test, exámenes, cuestionarios o encuestas dispone, en principio, de una lista de datos. Si se han observado pocos valores es posible que la simple inspección visual de los mismos sea suficiente para poder describir el fenómeno estudiado. Pero esto no es nada frecuente.

Si queremos, por ejemplo, después de un examen saber cuál es la puntuación del estudiante típico, cuál es el rango en que varían las puntuaciones, si los estudiantes se agrupan en las posiciones extremas o en las centrales o están dispersos, entonces tendremos que poner en orden nuestros datos de forma que podamos interpretarlos.

Un instrumento para conseguir esta ordenación de los datos es lo que llamamos distribución de frecuencias, que además de ésta función debe cumplir otras dos más: ofrecer la información necesaria para hacer representaciones gráficas y facilitar los cálculos para obtener los estadísticos muestrales.

Una distribución de frecuencias es, según Hays (1988), una representación de la relación entre un conjunto de medidas o clases de medidas mutuamente exclusivas y exhaustivas y la frecuencia de cada una de ellas.

Para definir el término frecuencia que aparece, a su vez, en esta definición vamos a establecer una notación: la letra X mayúscula representará a la variable con la que estamos trabajando. La letra X mayúscula con subíndices, X_1 X_2 X_3, servirá para representar un valor concreto de la variable X en el sujeto 1,2,3,...Cuando queramos referirnos a un valor concreto cualquiera de la variable X escribiremos X_i. El número de elementos que componen la muestra será n.

Se llama frecuencia de un valor X_i, y se simboliza por f_i al número de veces que se repite el valor X_i en la muestra.

Ahora vamos a seguir los pasos para la construcción de una tabla de distribución de frecuencias con un ejemplo sencillo.

Supongamos que un profesor pasa a sus alumnos una encuesta en la que, entre otras cosas, se les pregunta por el número de hermanos. Las respuestas de sus treinta alumnos son: 1, 2, 1, 1, 3, 2, 1, 2, 2, 3, 1, 1, 1, 2, 1, 2, 2, 1, 1, 4, 4, 2, 2, 3, 4, 3, 1, 3, 1, 1.

Para construir la tabla de distribución de frecuencias se inspeccionan en primer lugar los valores que toma la variable. En este caso se trata de una variable discreta que sólo toma los valores 1, 2, 3 y 4. En segundo lugar se cuenta cuántas veces aparece cada uno de ellos. Estos datos se colocan en una tabla de la forma siguiente:

Valores de la variable X_i	Frecuencia f_i
4	3
3	5
2	9
1	13
	n =30

Tabla 1: Distribución de frecuencias de la variable "Número de hermanos".

Una vez construida esta tabla y a pesar de su simplicidad, ya podemos extraer algunas conclusiones, por ejemplo que las familias con un único hijo son las más frecuentes. Si sumamos la frecuencia de éstas y la de las familias de dos hijos, 13 + 9 son 22 familias, que en tanto por ciento sobre 30 representan el 73,3% de la muestra.

Este es un ejemplo de tabla de distribución de frecuencias muy simple debido fundamentalmente a que la variable sólo toma cuatro valores diferentes.

Vamos a poner otro ejemplo algo más complejo: un profesor pasa un test de hábitos de estudio a sus treinta alumnos, los resultados son los siguientes: 37, 72, 71, 65, 54, 78, 85, 42, 49, 63, 61, 32, 51, 33, 77, 93, 85, 83, 63, 55, 58, 46, 57, 73, 73, 68, 73, 91, 75, 77.

El valor más pequeño es 32 y el mayor 93. Si construyésemos una tabla de distribución de frecuencias como la anterior tendríamos una lista demasiado extensa (62 números) y muchas de las frecuencias serían cero. En estos casos se recurre a lo que se denomina la agrupación en intervalos de clase, que consiste en formar grupos de valores consecutivos de la variable y poner cada uno de estos grupos en cada fila, en lugar de poner una sola puntuación.

Para agrupar las puntuaciones de la variable se suelen establecer estas dos reglas:

1º. Son preferibles los intervalos de clase que contengan 1, 2, 3, 5, 10 ó 20 unidades de la escala.

2º. El número de intervalos o grupos debe variar entre 10 y 20.

En nuestro ejemplo, la variable toma valores entre 32 y 93, es decir, su rango ó amplitud total es 93-32+1=62. Nos fijamos en la regla nº 2 y dividimos 62/10=6,2 y 62/20=3,1. Esto quiere decir que si agrupamos las puntuaciones de 6 en 6 tendremos 10 intervalos, si las agrupamos de 3 en 3 tendremos alrededor de 20 intervalos. Atendiendo a la regla nº 1, decidimos hacer intervalos de clase de amplitud 5.

La siguiente pregunta es ¿a partir de qué número empezamos a contar en la escala?. A partir de un número que sea múltiplo del tamaño de los intervalos de clase y que se aproxime lo más posible a la medida observada menor. En nuestro ejemplo 32 es el valor más pequeño y 30 el múltiplo de 5 más cercano, así que el primer intervalo de clase contendrá las puntuaciones 30, 31, 32, 33 y 34 y el último 90, 91, 92, 93 y 94. La tabla de distribución de frecuencias agrupada sería la siguiente:

Intervalos de clase X_i	Frecuencia del intervalo f_i
90-94	2
85-89	2
80-84	1
75-79	4
70-74	5
65-69	2
60-64	3
55-59	3
50-54	2
45-49	2
40-44	1
35-39	1
30-34	2
	n = 30

Tabla 2: Distribución de frecuencias con los datos agrupados en intervalos de clase de la variable puntuaciones obtenidas en un test de hábitos de estudio.

Según esta tabla de distribución de frecuencias agrupadas, la variable puntuación del test de hábitos de estudio no puede tomar valores entre 44 y 45 o entre 59 y 60. Aunque en la práctica esto es así porque el test usado para medir hábitos de estudio no tiene la precisión suficiente para obtener valores como 44,5 o 59,8, en teoría tenemos que considerar esta variable como continua en el intervalo, es decir, puede tomar cualquier valor entre 30 y 94. Así pues los límites exactos del intervalo 30-34 son 29,5-34,5, los del intervalo 35-39 son 34,5-39,5 y así sucesivamente de forma que el límite superior exacto de un intervalo coincida con el límite inferior exacto del siguiente. Por

otra parte, a los límites de los intervalos que aparecen en la Tabla 2, les llamaremos límites informados.

A partir de los límites informados o de los límites exactos se puede definir el punto medio del intervalo como el punto que resulta de la suma del extremo superior y el extremo inferior dividida por dos, es decir, como su media. Por ejemplo, el punto medio del intervalo 60-64 es el punto 62, resultado de (60 + 64) / 2 ó de (59,5 + 64,5) / 2.

La amplitud del intervalo se define como la diferencia entre el límite superior exacto y el límite inferior exacto.

En este momento volveremos a escribir nuestra tabla de distribución de frecuencias añadiendo los límites exactos de los intervalos y sus puntos medios, para usarla más adelante en las representaciones gráficas.

Intervalos de clase X_i	Límites exactos	Punto medio	Frecuencia f_i
90-94	89,5-94,5	92	2
85-89	84,5-89,5	87	2
80-84	79,5-84,5	82	1
75-79	74,5-79,5	77	4
70-74	69,5-74,5	72	5
65-69	64,5-69,5	67	2
60-64	59,5-64,5	62	3
55-59	54,5-59,5	57	3
50-54	49,5-54,5	52	2
45-49	44,5-49,5	47	2
40-44	39,5-44,5	42	1
35-39	34,5-39,5	37	1
30-34	29,5-34,5	32	2
			n = 30

Tabla 3: Intervalos de clase, límites exactos, puntos medios y frecuencias de las puntuaciones obtenidas en el test de hábitos de estudio.

Una vez construida la tabla de distribución de frecuencias, a la que hemos añadido los límites exactos de los intervalos y sus puntos medios, estamos en condiciones de hacer algunas representaciones gráficas que nos ayudarán a interpretar la situación de los alumnos en cuanto al test de hábitos de estudio.

Para hacer las representaciones gráficas de las tablas de distribución de frecuencias podemos considerar dos situaciones distintas:

1ª. Las observaciones dentro de un intervalo de clase están distribuidas uniformemente entre sus límites exactos.
2ª. Las observaciones dentro de un intervalo de clase están concentradas en su punto medio.

Si nos encontramos en la primera situación haremos un histograma o diagrama de columnas. En el eje de abcisas se representan los límites exactos de los intervalos de clase y en el eje de ordenadas la frecuencia de cada intervalo. Sobre cada uno de estos intervalos se dibuja un rectángulo cuya base está delimitada por los límites exactos y su altura es la frecuencia de ese intervalo.

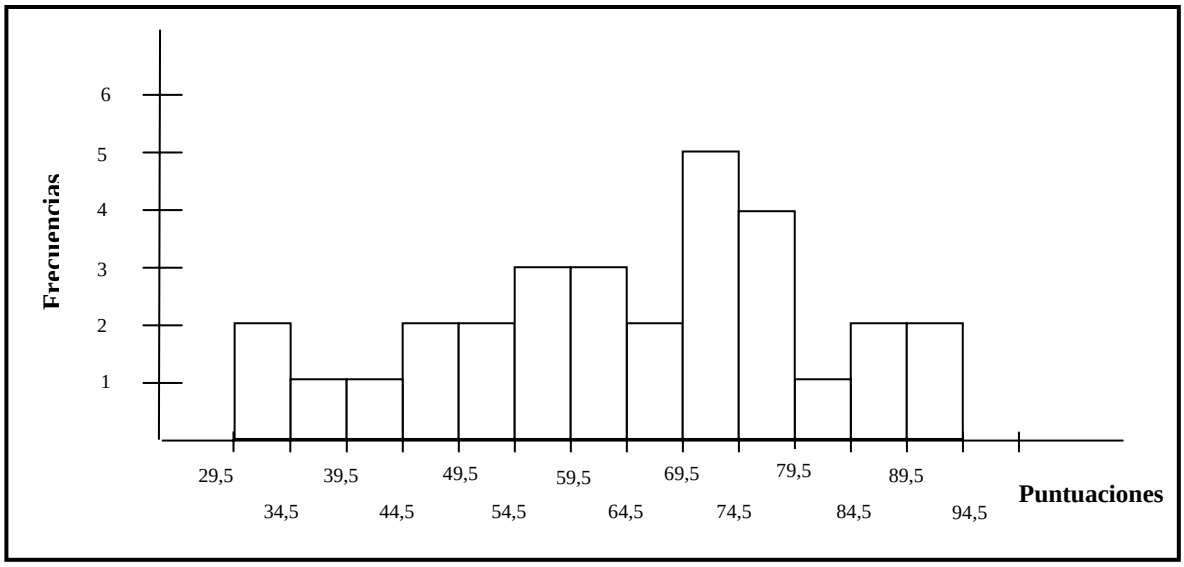

Figura 2: Histograma de los datos de la Tabla 3.

Si nos encontramos en la segunda situación, es decir, si consideramos que las observaciones dentro de cada intervalo se concentran en su punto medio, construiremos un polígono de frecuencias. En el eje de abcisas representaremos los puntos medios de cada intervalo y en el de ordenadas la frecuencia de cada intervalo. Uniendo estos puntos de forma consecutiva mediante segmentos de recta, obtendremos el polígono de frecuencias.

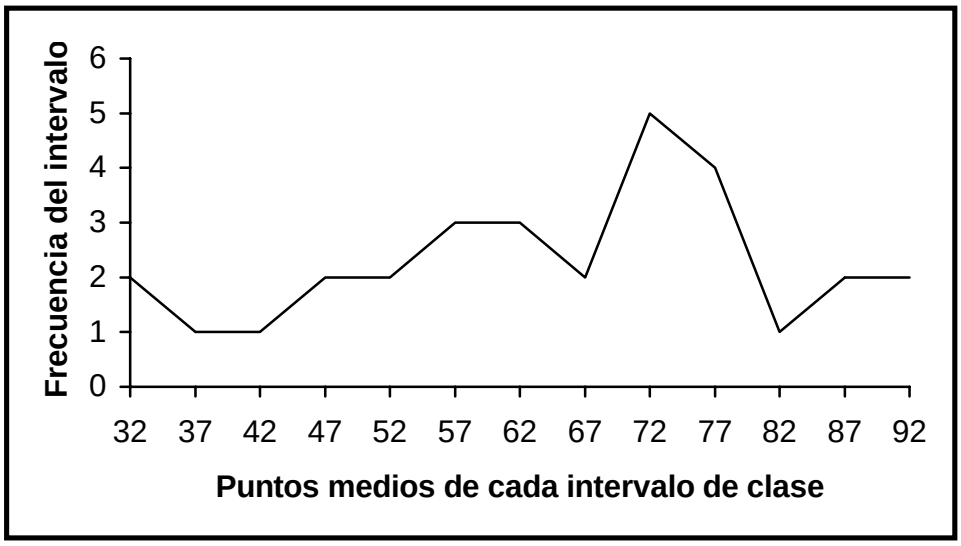

Figura 3: Polígono de frecuencias de los datos de la Tabla 3.

Es evidente que el histograma y el polígono de frecuencias ofrecen una imagen muy similar de la distribución de frecuencias de los datos. Esto es así porque se puede

construir el segundo a partir del primero sin más que unir los puntos medios de las bases superiores de los rectángulos del histograma.

Ahora vamos a hacer algunas consideraciones respecto a las dos normas para la construcción de tablas de distribuciones de frecuencia. En primer lugar queda claro que, partiendo de los mismos datos y teniendo en cuenta las dos reglas citadas anteriormente, se pueden hacer muchas tablas distintas, todas ellas igualmente válidas, es decir, que cumplen los requisitos para los cuales se han diseñado: la ordenación de una manera clara y sistemática de nuestros datos. En segundo lugar, tenemos que tener en cuenta que se nos pueden presentar situaciones en las que sea difícil aplicar estas reglas sin perder gran parte de la información. Por ejemplo, si estamos trabajando con la variable "ingresos mensuales", tendremos una mayoría de valores concentrados en torno a valores centrales y unos pocos que se desvían de éstos mucho por arriba. Si en un caso así se hiciera un número de intervalos en torno a diez, y de amplitud constante, la inmensa mayoría de los datos estarían concentrados en uno o dos intervalos. Para evitar eso se utiliza lo que se denomina intervalos abiertos, en los cuales no se considera límite superior o límite inferior. Por ejemplo, el primer intervalo podría ser "menos de 50.000" y el último "más de 500.000" .

2.- Distribución de frecuencias acumuladas, polígono de frecuencias acumuladas y polígono de porcentajes de frecuencia acumulada

Hasta el momento hemos presentado una de las formas de representación y tabulación de datos, ahora haremos referencia a la distribución de frecuencias acumuladas que se usa cuando se tiene interés en el número de observaciones que se sitúan por debajo de un cierto punto de la escala de medición.

Definiremos frecuencia acumulada de un intervalo de clase como el número de casos u observaciones dentro de dicho intervalo más todos aquellos contenidos en intervalos inferiores en la escala.

La frecuencia acumulada se calcula de forma inmediata a partir de la frecuencia de cada intervalo de clase. Para el primero de éstos ambas frecuencias coinciden, para los siguientes, la frecuencia acumulada es igual a su frecuencia más la acumulada del intervalo anterior. Así, la frecuencia acumulada del último intervalo será igual al número de observaciones de la distribución.

Cuando trabajamos con variables distintas o con la misma variable medida en muestras de distinto tamaño resulta difícil comparar las frecuencias absolutas. Una forma de resolver este problema es calcular los porcentajes de frecuencias de cada intervalo de clase, ya que así compararemos dos distribuciones con cien observaciones. De la misma forma se calculan porcentajes de frecuencias acumuladas.

Volvamos al ejemplo de los datos del test de hábitos de estudio, para construir la tabla de la distribución de frecuencias acumuladas calculando los porcentajes de frecuencias acumuladas

Intervalos de clase X_i	Límites exactos	Frecuencia f_i	Frecuencia acumulada	Porcentaje de frec. acumulada
90-94	89,5-94,5	2	30	100,00
85-89	84,5-89,5	2	28	93,33
80-84	79,5-84,5	1	26	86,67
75-79	74,5-79,5	4	25	83,33
70-74	69,5-74,5	5	21	70,00
65-69	64,5-69,5	2	16	53,33
60-64	59,5-64,5	3	14	46,67
55-59	54,5-59,5	3	11	36,67
50-54	49,5-54,5	2	8	26,67
45-49	44,5-49,5	2	6	20,00
40-44	39,5-44,5	1	4	13,33
35-39	34,5-39,5	1	3	10,00
30-34	29,5-34,5	2	2	6,67

Tabla 4. Distribución de frecuencias acumuladas y porcentaje de frecuencias acumuladas de las puntuaciones en el test de hábitos de estudio.

A partir de esta tabla podemos representar gráficamente el polígono de frecuencias acumuladas y el polígono de porcentajes de frecuencias acumuladas, representando en el eje de abcisas los límites superiores exactos de los intervalos de clase y en el de ordenadas las correspondientes frecuencias acumuladas o porcentajes de frecuencias acumuladas respectivamente. Estos dos polígonos tienen exactamente la misma forma, difieren únicamente en la escala del eje de ordenadas. Estos gráficos nos permiten conocer cuántas observaciones o qué porcentaje se sitúan por debajo de un intervalo de clase.

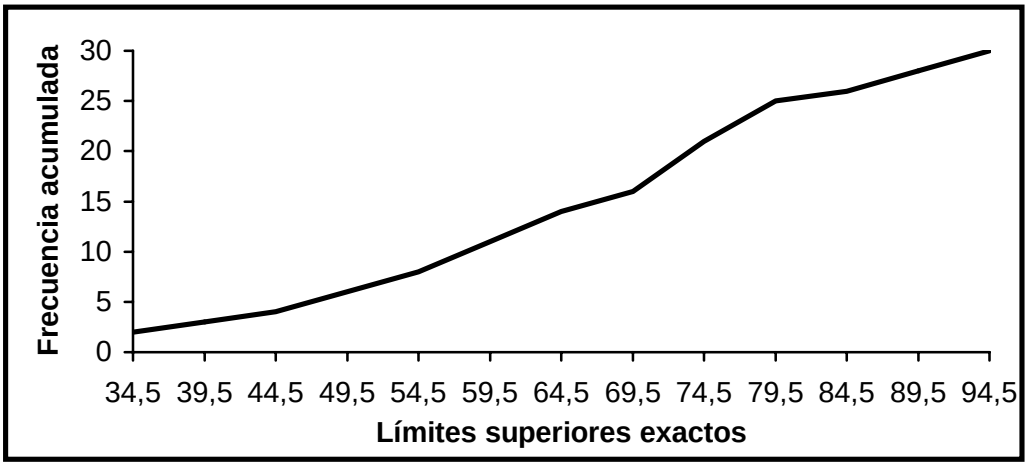

Figura 4: Polígono de frecuencias acumuladas a partir de los datos de la Tabla 4.

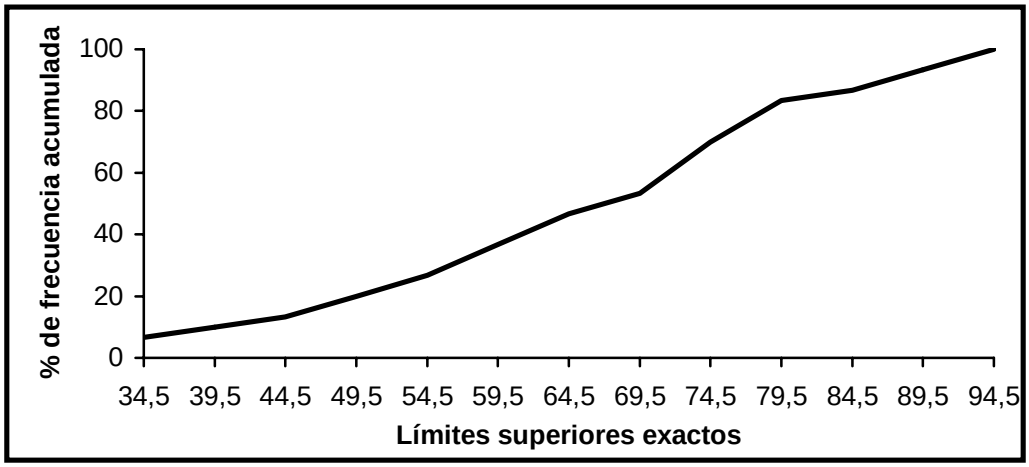

Figura 5: Polígono de porcentajes de frecuencia acumulada a partir de los datos de la Tabla 4.

3.- Otras representaciones gráficas

De entre las distintas representaciones gráficas que se pueden hacer con nuestros datos haremos sólo referencia a los pictogramas y los perfiles por su uso más extendido.

Un pictograma es una representación gráfica de una variable en forma de círculo que está dividido en tantos sectores como valores distintos tome la variable. Además la superficie de los mismos es proporcional a la frecuencia de cada modalidad de la variable. En la figura aparece el pictograma de la variable "Estudios de los padres". Además de las modalidades de la misma hemos añadido el tanto por ciento que representa cada una. En ocasiones, cuando se quiere destacar una de las secciones en particular se separa del resto para captar la atención del lector en esa modalidad particular.

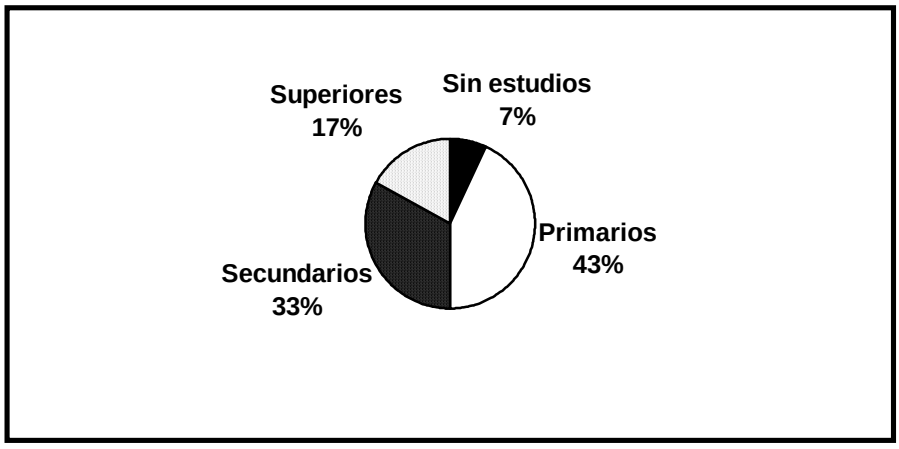

Figura 6: Pictograma de la variable "Estudios de los padres".

Los perfiles se usan bastante en informes psicopedagógicos o de rendimiento. En el eje de ordenadas se representan las puntuaciones alcanzadas en distintas variables o parámetros, todos ellos medidos con la misma unidad. Por ejemplo: las calificaciones de un alumno en cuatro asignaturas distintas medidas de uno a diez, o las medias de la asignatura de Matemáticas de 1º de ESO en los grupos A, B, C y D. Y en el eje de abcisas se coloca una marca por cada sujeto, grupo o variable medida. En los ejemplos, un marca por cada asignatura o una marca para cada uno de los grupos. En la figura

aparece el perfil correspondiente a las calificaciones de un alumno en cuatro asignaturas.

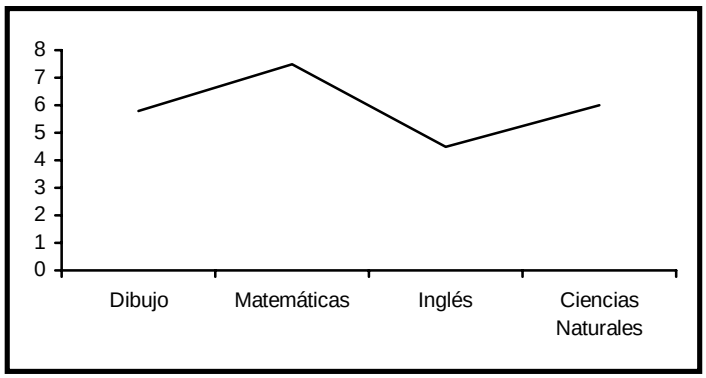

Figura 7: Perfil de las calificaciones de un alumno.

4.- El rango del percentil

Supongamos que el padre de uno de nuestros alumnos acude a la tutoría para interesarse por las notas de su hijo. Nuestra respuesta puede ser: en Matemáticas obtuvo un 5.5, en Literatura un 6.8 y en Filosofía un 8. A primera vista puede ser una información valiosa, pero si queremos sacarle un poco más de jugo que el aparente, enseguida nos damos cuenta de que nos faltan bastantes datos. En general, para poder interpretar el significado de una puntuación es necesario hacerlo en términos relativos y con respecto a un grupo de referencia. En nuestro caso necesitaríamos saber dónde está nuestro alumno en relación a los alumnos de su clase.

Para hacer estas valoraciones relativas se utilizan las llamadas medidas de posición, que son índices diseñados para revelar la situación de una puntuación con respecto a un grupo, utilizando a éste como marco de referencia. Una de las medidas de posición más utilizadas son los percentiles también llamados centiles ó el rango del percentil.

Los centiles o percentiles son 99 valores de la variable que dividen a la distribución en 100 secciones, cada una de ellas conteniendo a la centésima parte de las observaciones. Una puntuación X corresponde al rango del percentil K, cuando el K% de las observaciones se sitúan por debajo de X.

Una forma de calcular aproximadamente las puntuaciones a las que corresponde tal o cual percentil es a través del polígono de porcentajes de frecuencias acumuladas. Si queremos calcular a qué puntuación corresponde el percentil 50, trazamos a la altura del 50 en el eje de ordenadas una línea paralela al eje de abcisas hasta interceptar el polígono. En ese punto trazamos una línea paralela al eje de ordenadas hasta interceptar con el eje de abcisas y ese punto es precisamente el que se corresponde al percentil 50. Si seguimos estos pasos en la Figura 5, que corresponde al polígono de porcentajes de frecuencias acumuladas de los datos obtenidos del test de hábitos de estudio, el percentil 50 correspondería aproximadamente a la puntuación 67.

La fórmula para obtener el percentil K es :

$$P_k = L_i + \left[\left(\frac{I}{n_i} \right) \cdot \left(\left(\frac{k \cdot n}{100} \right) - n_a \right) \right]$$

donde:

- P_k es la puntuación correspondiente al percentil k.
- L_i es el límite inferior exacto del intervalo crítico (aquel que acumula al menos el k% de la frecuencia acumulada).
- I es la amplitud de los intervalos.
- n_i es la frecuencia del intervalo crítico.
- k es el porcentaje de observaciones inferiores a P_k.
- n es el número de observaciones hechas.
- n_a es la frecuencia acumulada hasta L_i .

Ahora vamos a poner un ejemplo para ilustrar el uso de esta fórmula. Vamos a calcular en primer lugar la puntuación a la que corresponde el percentil 40. Tendremos que determinar el intervalo que acumula al menos el 40% de las observaciones. Miramos la Tabla 4, distribución de frecuencias acumuladas, en la columna de porcentaje de frecuencia acumulada. En nuestro caso se trata del intervalo cuyos límites exactos son 59,5-64,5 cuyo porcentaje de frecuencia acumulada es 46,67%, y una vez localizado este intervalo ya tenemos todos los datos requeridos en la fórmula: L_i = 59,5, I = 64,5 – 59,5 = 5, n_i = 3, k = 40, n=30, n_a = 11 y, por tanto

P_k= 59,5 + (5 / 3) · (12 - 11) = 61,17

La puntuación 61,17 corresponde al percentil 40, es decir el 40% de las observaciones se sitúan por debajo de 61,17. Aprovechamos este resultado para indicar que la puntuación que corresponde a un determinado percentil no tiene por qué coincidir con una puntuación observada como ocurre en este ejemplo. Ninguno de los alumnos del ejemplo obtenía en el test de hábitos de estudio una puntuación de 61,17.

Utilizaremos en forma inversa la fórmula para calcular a qué percentil corresponde una determinada puntuación. Un alumno ha obtenido la puntuación 85 y nos pregunta a qué percentil se corresponde o más bien nos pregunta si es una puntuación alta o baja en relación con sus compañeros de clase. Empleamos la misma fórmula aunque en éste caso la incógnita es k. P_k es 85 , el intervalo crítico es aquél que contiene la puntuación 85, cuyos límites son 84,5-89,5. Por tanto, L_i = 84,5, I = 5, n_i = 2, n=30, n_a = 26. Sustituyendo en la fórmula tenemos:

85 = 84,5 + (5 / 2) · (((k ·30) / 100) - 26),
85 – 84,5 = 2.5 · ((k · 0,3) - 26),
0,5 / 2,5 = (k · 0,3) - 26,
0,2 + 26 = k · 0,3
k = 26,2 / 0,3 = 87,33

La puntuación de nuestro alumno se corresponde con el percentil 87,33, el 87,33% de los resultados del test de hábitos de estudio se encuentran por debajo de su puntuación, lo cual quiere decir que su puntuación destaca dentro de la clase.

Nos hemos referido al rango del percentil como una medida de posición que nos permite comparar observaciones de una variable respecto de las observaciones de la misma en un grupo o comparar variables distintas en un mismo grupo. Pero como tal medida de posición tiene también sus inconvenientes. El más importante es el que se deriva del hecho de que se está utilizando una escala ordinal: las mismas diferencias en percentiles no se corresponden con diferencias en la puntuación de la variable. Las diferencias en las puntuaciones correspondientes a los percentiles 55 y 56 no tienen por qué ser iguales a las diferencias entre las puntuaciones que corresponden a los percentiles 93 y 94. Generalmente, las distancias entre centiles intermedios suelen ser menores que las diferencias entre centiles extremos, y esto es así porque normalmente se obtienen con más frecuencia puntuaciones intermedias de las variables y los valores más extremos son más infrecuentes.

Una vez definidos los percentiles podemos hacer referencia a otras medidas de posición que se obtienen directamente de ellos: los cuartiles y los deciles. Los cuartiles son tres y se denotan por Q_1, Q_2 y Q_3. Se definen como los tres valores de la variable que dividen a la distribución en cuatro partes, cada una contiendo al 25 por cien de las observaciones. Por lo tanto el primer cuartil coincide con el percentil 25, el segundo con el percentil 50 y el tercero con el percentil 75.

Los deciles se representan por D_k, donde k representa el número del decil al que se refiere, y son nueve puntuaciones que dividen a la variable en 10 partes cada una conteniendo el 10 por 100 de las observaciones. El primer decil corresponde al percentil 10, el segundo al percentil 20 y así sucesivamente.

5.- Características generales de una distribución de frecuencias

Hasta este momento hemos visto cómo construir tablas y representar gráficamente un conjunto de datos, pero estas técnicas todavía no son suficientes para hacer comparaciones entre distintas distribuciones de frecuencias. Para ello es necesario definir algunas características de las distribuciones de frecuencias que llamamos: tendencia central, variabilidad o dispersión, sesgo y curtosis. Todas ellas tienen sus correspondientes medidas, es decir, sus indicadores que obtenemos mediante una serie de cálculos a partir de los datos de una tabla de distribución de frecuencias.

Supongamos que el jefe de estudios del centro nos pregunta: ¿Cuál es el rendimiento de la clase A en tu asignatura? ¿Cuánto tiempo han empleado tus alumnos en recorrer el circuito?. Le podríamos responder presentándole las tablas de distribuciones de frecuencias o incluso las listas con los nombres de los alumnos, sus calificaciones y tiempos, pero casi con toda seguridad le costaría un gran esfuerzo entresacar de esos datos respuestas precisas. Otra forma de responder a las preguntas más claramente sería calcular algunas medidas de tendencia central de esas distribuciones, que precisamente son representaciones del valor "típico" o "promedio" de la variable, que se refieren al centro de la distribución de frecuencias, a la puntuación que representa a todas las demás. Las medidas de tendencia central más utilizadas son la media, la mediana y la

moda. Por su importancia, dedicaremos a su cálculo, su significado y sus propiedades una parte del capítulo siguiente.

La variabilidad o dispersión se refiere al grado de concentración de las observaciones en torno al promedio. Una distribución de frecuencias será homogénea o poco variable si los datos difieren poco entre sí y ,por tanto, se agrupan en torno a su promedio. Por el contrario será heterogénea o muy variable si los datos se dispersan mucho respecto al promedio. Esta propiedad es independiente de la tendencia central, es decir, dos distribuciones pueden tener la misma media y distinta variabilidad y viceversa. Precisamente, esta independencia es la causa de la importancia de la variabilidad, porque si prescindiésemos de ella, podríamos confundir, por tener el mismo rendimiento medio, una clase con alumnos de rendimiento medio muy similar con otra que tuviera alumnos de rendimiento máximo y alumnos de rendimiento mínimo. Las medidas de variabilidad o dispersión más frecuentes son: las desviación típica, desviación media, la varianza, el rango y la amplitud semiintercuartil. Al igual que con la tendencia central, al cálculo y propiedades de estas medidas dedicaremos parte del capítulo siguiente.

En la Figura 8 aparecen las representaciones de las distribuciones de frecuencia de tres grupos A, B y C. Los grupos A y B son iguales en cuanto a tendencia central y diferentes en cuanto a variabilidad, el grupo B es más variable que A. Los grupos A y C son iguales en cuanto a dispersión y difieren en su tendencia central: la de A es menor que la de C.

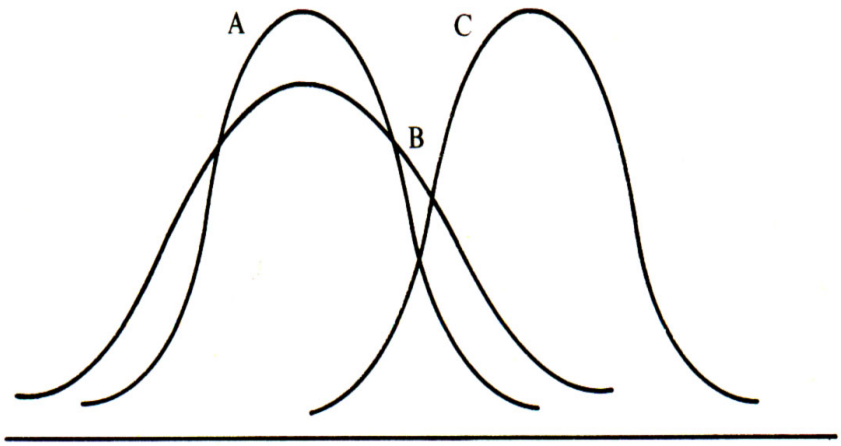

Figura 8: Ejemplos de tres distribuciones en las que A y B tienen tendencias centrales similares, y menores que la C, mientras que con respecto a la variabilidad la de B es mayor que las otras dos.

El sesgo o asimetría se refiere al grado en que los datos tienden a concentrarse en los valores centrales, en los valores inferiores al promedio, o en los valores superiores a éste. También podemos decir que hace referencia al grado en que los datos se reparten equilibradamente por encima y por debajo de la tendencia central. Una distribución será simétrica cuando, al dividirla en dos a la altura de la media, las dos mitades se superponen. Una distribución tiene asimetría positiva cuando la mayor concentración de puntuaciones se produce en la parte baja de la escala mientras que algunas puntuaciones

son altas. Una distribución tiene asimetría negativa cuando la mayor parte de las observaciones se sitúan en la parte alta de la escala mientras que se produce alguna observación en la parte baja. Un ejemplo: si ponemos a nuestros alumnos un examen muy fácil, la distribución de frecuencias de sus calificaciones tendrá sesgo negativo puesto que la mayoría de los alumnos obtendrían calificaciones altas. Si por el contrario el examen es difícil, estaremos ante una distribución con sesgo positivo puesto que la mayoría de los alumnos obtendrían notas bajas y sólo algunos destacarían con notas altas. Si el examen es de dificultad media, lo más probable es que la distribución sea simétrica. Para esta característica no vamos a estudiar ningún índice porque normalmente los cálculos son laboriosos, más bien podemos dar un criterio aún cuando la podemos apreciar mediante la inspección visual del polígono de frecuencias. El criterio tiene que ver con las diferencias entre cuartiles y es el siguiente:

Si $(Q_3 - Q_2) > (Q_2 - Q_1)$ entonces la distribución tiene sesgo positivo.
Si.$(Q_3 - Q_2) < (Q_2 - Q_1)$ tiene sesgo negativo.
Si $(Q_3 - Q_2) = (Q_2 - Q_1)$ la distribución será simétrica .

Respecto a la inspección gráfica, en la Figura 9 aparecen las gráficas de tres grupos A, B y C. La distribución A es simétrica, la B tiene sesgo positivo y la C negativo.

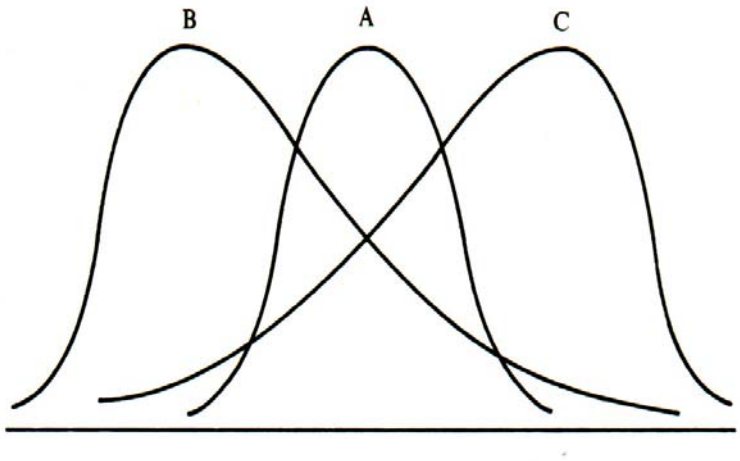

Figura 9: Ejemplos de distribuciones con distinto tipo de sesgo. La A es simétrica, la B asimétrica positiva y la C asimétrica negativa.

La curtosis se usa para saber cómo es de escarpado o plano un polígono de frecuencias. El concepto de curtosis sólo se aplica a distribuciones unimodales (distribuciones que tienen un único “pico”) y se refiere al empinamiento de la curva en la proximidad de la moda. Generalmente el grado de curtosis de una distribución se compara con un modelo de distribución que estudiaremos más adelante que es la llamada campana de Gauss o distribución normal. Así, las distribuciones que tienen el mismo grado de apuntamiento que la normal se llaman mesocúrticas. Las distribuciones que tienen mayor grado de apuntamiento que la normal se llaman leptocúrticas y las que lo tienen menor platicúrticas. Los índices empleados habitualmente para calcular la

curtosis son demasiado complicados, comparados con su utilización, por lo que en estas notas no haremos referencia a ellos.

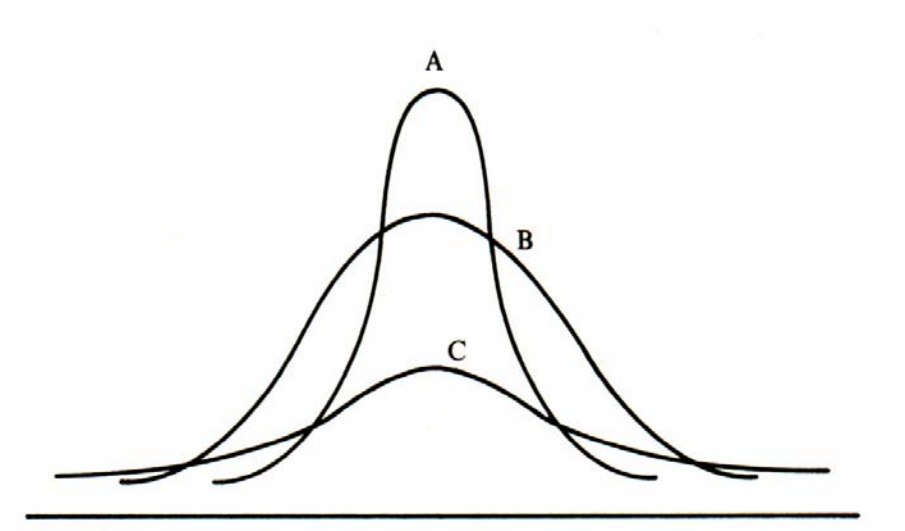

Figura 10: Ejemplos de distribuciones con distintos tipos de curtosis.
La A es leptocúrtica, la B mesocúrtica y la C platicúrtica.

Capitulo 3. Medidas de tendencia central

1.- La media.

La media de una variable se define como la suma de todos los valores observados dividida por el número de ellos. Se denota por la misma letra que la variable con una barra horizontal encima. Si tenemos n valores de la variable X su media se calcula como:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Por ejemplo, las notas obtenidas por una clase de 20 alumnos en un examen de Historia y ordenadas de menor a mayor son : 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 7, 7, 8 y 9. Su media se calcula :

(1+2+2+3+3+3+4+4+4+5+5+5+5+5+6+6+7+7+8+9) / 20 = 94 / 20 = 4,7

También podemos calcular la media a partir de la distribución de frecuencias:

$$\overline{X} = \frac{\sum_{i=1}^{k} X_i \cdot f_i}{\sum_{i=1}^{k} f_i}$$

dónde k representa el número de valores distintos que toma la variable y f_i la frecuencia de la puntuación i. Naturalmente, la suma de las frecuencias de las puntuaciones, el denominador de la fórmula anterior, tiene que ser igual al número de datos observados. Es decir:

$$\sum_{i=1}^{k} f_i = n$$

Veamos ahora cómo podemos usar la tabla de distribución de frecuencias, empleando los datos del ejemplo, para aplicar la fórmula anterior del cálculo de la media. Como en el numerador aparece la suma de los productos de cada puntuación por su frecuencia, añadimos una columna más a la tabla de distribución de frecuencias en la que escribimos precisamente cada uno de estos sumandos. La media será el cociente entre la suma de la tercera columna y la suma de la segunda de la siguiente Tabla 5. El resultado para la media es:

$$\overline{X} = \frac{94}{20} = 4{,}7$$

X_i	f_i	$X_i \cdot f_i$
9	1	9
8	1	8
7	2	14
6	2	12
5	5	25
4	3	12
3	3	9
2	2	4
1	1	1
Total	20	94

Tabla 5: Tabla de distribución de frecuencias para el cálculo de la media

Si estamos ante una distribución de frecuencias agrupadas y por tanto no disponemos de los datos observados, para calcular la media en lugar de los datos observados trabajaremos con los puntos medios de los intervalos de clase y la frecuencia de cada uno de ellos. Por tanto, si denotamos como Xm_i el punto medio del intervalo i-ésimo y f_i la frecuencia del mismo, calcularemos la media como:

$$\overline{X} = \frac{\sum_{i=1}^{k} Xm_i \cdot f_i}{\sum_{i=1}^{k} f_i}$$

Hay que tener en cuenta que en las otras fórmulas el sumatorio se extendía a lo largo de todas las puntuaciones observadas y a lo largo de las puntuaciones distintas observadas, respectivamente, en este caso el sumatorio tiene tantos sumandos como intervalos de clase. Es decir, en esta expresión k representa el número de intervalos de clase.

Para ilustrar el uso de esta fórmula, valiéndonos de las tablas de distribución de frecuencia, retomamos la Tabla 3 de las puntuaciones de un test de hábitos de estudio pasado a 30 alumnos. A esta tabla añadimos una columna que es el resultado de multiplicar cada punto medio del intervalo de clase por su frecuencia. Así, la media será el resultado de dividir el total de esta columna entre el total de las frecuencias, tal como aparece en la Tabla 6:

$$\overline{X} = \frac{1940}{30} = 64{,}67$$

X_i	Xm_i	f_i	$Xm_i \cdot f_i$
90-94	92	2	184
85-89	87	2	174
80-84	82	1	82
75-79	77	4	308
70-74	72	5	360
65-69	67	2	134
60-64	62	3	186
55-59	57	3	171
50-54	52	2	104
45-49	47	2	94
40-44	42	1	42
35-39	37	1	37
30-34	32	2	64
	Total	30	1940

Tabla 6: Distribución de frecuencias agrupadas en intervalos de clase para el cálculo de la media

Estos dos ejemplos nos sirven también para hacer una observación: nótese que la media no tiene por qué coincidir con una puntuación observada, aunque se haya definido como una puntuación que "representa a todas".

Otra interpretación de la media, desde el punto de vista físico, consiste en considerarla el centro de gravedad de las puntuaciones. Si a lo largo de una barra pusiésemos una unidad de peso sobre cada valor observado y tantas unidades como veces se haya observado el valor, entonces esa barra sólo estaría en equilibrio si el fulcro estuviese colocado a la altura de la media.

Este hecho, que constituye una de las principales propiedades de la media, se puede expresar matemáticamente diciendo que la suma de las diferencias de n puntuaciones respecto de su media es igual a cero. En efecto:

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)=\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\overline{X}=\sum_{i=1}^{n}X_i-n\cdot\overline{X}=\sum_{i=1}^{n}X_i-n\cdot\frac{\sum_{i=1}^{n}X_i}{n}=\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}X_i=0$$

Si en lugar de sumar las diferencias de las puntuaciones respecto de la media sumamos sus cuadrados, entonces encontramos otra propiedad interesante de la media: que esta suma es menor que si tomásemos las diferencias respecto de otro valor cualquiera. Expresado en lenguaje matemático:

$$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2<\sum_{i=1}^{n}\left(X_i-C\right)^2$$

Siendo C una constante cualquiera distinta de la media.

Supongamos ahora que sabemos la media de las clases A, B y C de las puntuaciones de la última evaluación de Química y queremos saber la media de los alumnos de los tres grupos juntos. Podríamos calcularla a través de las puntuaciones directas pero podemos aprovechar la información que ya tenemos porque la media total sería:

$$\overline{X} = \frac{n_A \cdot \overline{X}_A + n_B \cdot \overline{X}_B + n_C \cdot \overline{X}_C}{n_A + n_B + n_C}$$

dónde n_A y X_A son, respectivamente, el número y la media de los alumnos de la clase A, n_B y X_B son, respectivamente, el número y la media de los alumnos de la clase B y n_C y X_C son, respectivamente, el número y la media de los alumnos de la clase C.

Para comprobar con un ejemplo las propiedades de la media imaginemos que las 10 notas de los alumnos de la clase son: 3, 3, 4, 5, 5, 5, 6, 7, 7 y 9. La media de estas notas es 5,4. Vamos a comprobar que la suma de las diferencias de cada puntuación respecto de su media es cero:

$2 \cdot (3 – 5,4) + (4 – 5,4) + 3 \cdot (5 – 5,4) + (6 – 5,4) + 2 \cdot (7 – 5,4) + (9 – 5,4)=$
$(2 \cdot (-2,4)) + (-1,4) + (3 \cdot (-0,4)) + 0,6 + (2 \cdot 1,6) + 3,6 =$
$-4,8 – 1,4 – 1,2 + 0,6 + 3,2 + 3,6 = 0$

Ahora comprobaremos que la suma de las diferencias al cuadrado respecto de la media es más pequeña que la suma de las diferencias al cuadrado respecto a otra constante:

$(2 \cdot (-2,4)^2) + (-1,4)^2 + (3 \cdot (-0,4)^2) + (0,6)^2 + (2 \cdot (1,6)^2) + (3,6)^2 =$
$(2 \cdot 5,76) + 1,96 + (3 \cdot 0,16) + 0,36 + (2 \cdot 2,56) + 12,96 = 32,4$

Esta es la suma de las diferencias al cuadrado respecto de la media. Ahora elegimos como constante el número 5 para calcular la suma de diferencias al cuadrado:

$2 \cdot (3 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + 2 \cdot (7 - 5)^2 + (9 - 5)^2 =$
$(2 \cdot 4) + 1 + 1 + (2 \cdot 4) + 16 =34.$

En este caso, 34 es mayor que 32,4, con lo que queda comprobada la segunda propiedad de la media. Hay que hacer notar que esta es una mera comprobación y no una demostración rigurosa.

2.- La mediana

La mediana es un punto de la escala de medida que divide a la distribución en dos partes iguales, es decir, la mitad de las puntuaciones son mayores que la mediana y la otra mitad son menores. Este índice se representa por *Md*.

A la hora de calcularla podemos encontrarnos frente a puntuaciones directas o, por el contrario, frente a una distribución de frecuencias agrupadas. Si estamos en el primer caso, a su vez nos podemos encontrar ante un número impar de observaciones o ante un número par. En primer lugar tenemos que ordenar las puntuaciones directas de

menor a mayor, si su número, n, es impar entonces la mediana es el lugar que ocupa la posición (n+1) / 2. Si n es par, la mediana es el punto medio entre los dos valores centrales, que en este caso son el valor n / 2 y (n / 2) + 1.

Cuando disponemos de una distribución de frecuencias agrupada en intervalos de clase, la mediana coincide con el percentil 50, puesto que justamente es aquella puntuación por debajo de la cuál se sitúan el 50 % de las observaciones y por tanto el otro 50% son mayores que ella. Por tanto, para calcularla usamos la fórmula de los percentiles haciendo k=50 lo que resulta:

$$Md = L_i + \left[\left(\frac{I}{n_i} \right) \cdot \left(\frac{n}{2} - n_a \right) \right]$$

donde:

- L_i es el límite inferior exacto del intervalo crítico.
- I es la amplitud de los intervalos.
- n_i es la frecuencia del intervalo crítico.
- n es el número de observaciones hechas.
- n_a es la frecuencia acumulada hasta L_i .

Si tomamos como ejemplo para calcular la mediana el citado en el capítulo 2 referido a las notas obtenidas por 20 alumnos en un examen de Historia, al ser un número par de observaciones que ya están ordenadas de menor a mayor, buscamos las que ocupan los lugares 20 / 2 = 10 y (20/ 2) + 1 = 11 que son en ambos casos un 5. El punto medio en este caso es (5 + 5) / 2 = 5. La mediana es la puntuación 5.

Supongamos ahora que queremos saber la mediana de las 11 puntuaciones siguientes: 7, 11, 6, 5, 7, 12, 9, 8, 10, 6 y 9. En primer lugar, las ordenamos de menor a mayor: 5, 6, 6, 7, 7, 8, 9, 9, 10, 11, 12. Ahora buscamos la que ocupa la posición (11 + 1) / 2 = 6 que en este caso es 9. La mediana de estas 11 puntuaciones es 9.

El resultado de un test de conocimientos previos sobre Geografía realizado por 200 alumnos queda reflejado en la siguiente tabla.

X_i	f_i	f_a
18-20	40	200
15-17	50	160
12-14	40	110
9-11	30	70
6-8	25	40
3-5	15	15

Tabla 7: Distribución de frecuencias agrupadas en intervalos y frecuencias acumuladas de las puntuaciones de un test de conocimientos previos sobre Geografía.

Determinamos la mediana de las puntuaciones usando la Tabla 7. El intervalo crítico será aquél que cuya frecuencia acumulada sea 100 ó más, puesto que el 50% de 200 es 100. Por tanto el intervalo crítico es el 12-14, su límite inferior exacto es 11,5, la amplitud es 3, la frecuencia es 40 y la frecuencia acumulada hasta ese intervalo es 70. Sustituyendo los valores en la fórmula:

$$Md = 11{,}5 + \left[\left(\frac{3}{40}\right)\cdot(100-70)\right] = 11{,}5 + \left(\frac{90}{40}\right) = 13{,}75$$

También podemos estimar gráficamente la mediana utilizando el polígono de porcentajes de frecuencias acumuladas. Para ello, se traza una paralela al eje de abcisas a la altura del 50% del de ordenadas hasta interceptar el polígono y, desde este punto, una paralela al eje de ordenadas hasta interceptar el de abcisas. Precisamente este punto de intersección será una estimación de la mediana porque el proceso mencionado anteriormente no es más que una "búsqueda gráfica" del valor de la abcisa al que le corresponde la ordenada 50. Dicho de otra forma, hemos buscado aquel punto que acumula el 50% de las observaciones, tiene por debajo de sí el 50% de las mismas y, por tanto, el mismo porcentaje por encima de sí. Y esta es la propiedad que define a la mediana. En la Figura 11 queda representado este proceso para las puntuaciones de la Tabla 7.

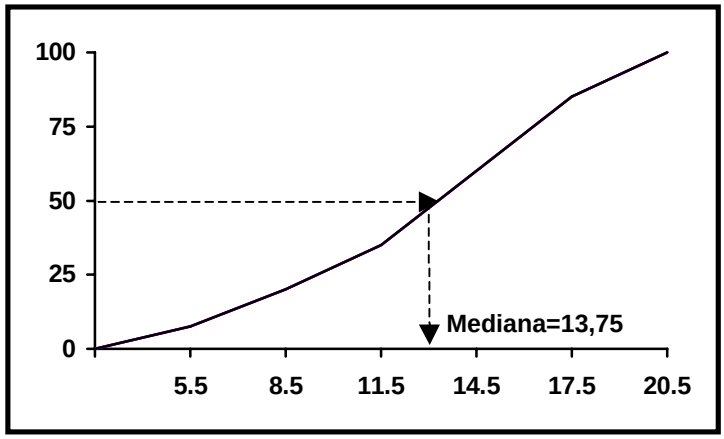

Figura 11: Estimación gráfica de la mediana para los datos de la Tabla 7.

3.- La moda

La moda, que se representa por *Mo,* es la medida de tendencia central más fácil de calcular ya que se define como el valor de la variable con mayor frecuencia. Sin embargo, no todas las distribuciones tienen moda, ni todas tienen una única moda, por lo que es necesario hacer referencia a distintos casos que se pueden presentar.

En primer lugar un caso que no ofrece ninguna duda: en el siguiente grupo de puntuaciones, 3, 4, 4, 5, 7, 9, 9, 9 y 10, la moda es 9. Sin embargo en este grupo: 1, 1, 2, 2, 3, 3, 4 y 4 no hay moda, se dice que la distribución es amodal porque todas las puntuaciones tienen la misma frecuencia.

Cuando dos puntuaciones adyacentes tienen la misma frecuencia y esta frecuencia común es mayor que cualquier otra puntuación, la moda es el promedio de las dos puntuaciones adyacentes. Por ejemplo, la moda de 1, 1, 2, 2, 2, 3, 3, 3, 5 y 6 es el promedio de 2 y 3, es decir, 2.5.

Si en un grupo de puntuaciones hay dos que no son adyacentes y tienen la misma frecuencia y esta frecuencia común es mayor que la de cualquier otra puntuación existen dos modas. En el conjunto 5, 5, 6, 6, 6, 7, 8, 8, 8, 9, 9, hay dos modas 6 y 8. En este caso la distribución de frecuencias se llama bimodal.

En general, podemos decir que la moda es el valor de la variable que se encuentra por debajo del pico más alto del polígono de frecuencias, con lo que queda definida para datos agrupados en intervalos de clase como el punto medio del intervalo de mayor frecuencia y además, se pueden aplicar las mismas normas anteriormente citadas respecto a las puntuaciones a los intervalos de clase. En el test de conocimientos previos de Geografía, cuyos resultados quedan reflejados en la Tabla 7, la moda corresponde a la puntuación 16, que es el punto medio del intervalo 15-17, ya que la frecuencia de ese intervalo es 50, mayor que la de cualquiera de los otros.

Cuando estamos ante conjuntos de datos muy numerosos se consideran distribuciones bimodales cuando presentan un polígono de frecuencias con dos picos, aún cuando las frecuencias correspondientes a cada uno de ellos no sean exactamente iguales. En todo caso, si se quiere hacer alguna distinción entre las modas, se suele hablar de moda absoluta, la que satisface la definición de moda, en contraposición de moda relativa, la que se corresponde con otros picos de la distribución de frecuencias menos elevados que la moda absoluta.

4.- Comparación entre medidas de tendencia central

Hasta aquí hemos expuesto los procedimientos de cálculo y las propiedades de las tres medidas de tendencia central que habitualmente usamos y nos queda hacer referencia a los criterios para elegir una u otra como representación de la puntuación general de una variable. Pues bien, en general preferiremos la media, primero porque se va a usar en cálculos posteriores con mucha frecuencia y segundo porque es la más estable. Es decir, de una muestra a otra varía menos que la mediana o la moda, está más cerca de la media de la población de lo que lo están la mediana y la moda de estos valores calculados en la población.

Elegiremos la mediana como medida de tendencia central cuando la distribución sea muy sesgada, cuando tenga valores muy extremos, ya que, en estos casos, la media se desplaza hacia las puntuaciones extremas y no así la mediana. La media se ve afectada por todos los valores de la variable. Si en una distribución cambiamos una puntuación, tendremos que volver a calcular la media mientras que la mediana sólo se ve afectada por los valores centrales. Por ejemplo en la serie 2, 3, 5, 6, 8, 8 y 19 la media es 7,28 y la mediana es 6. Si la puntuación 19 fuese fruto de un error tipográfico, y en realidad se tratase de un 9, entonces la media sería 5,85 y la mediana 6.

En el caso de encontrarnos ante una variable ordinal también elegiríamos la mediana como medida de tendencia central.

Y, por último, en caso de tener una distribución de frecuencias con intervalos de clase abiertos, es decir cuando el intervalo inferior o el superior carecen de algún límite ya que la media no se podría calcular al carecer estos intervalos de punto medio. A este tipo de distribuciones hicimos referencia en el capítulo 2, y el ejemplo de la variable "sueldo mensual" al que aludíamos es igualmente válido ahora. No podemos saber el punto medio del intervalo "menos de 50.000" ni el del "más de 500.000", por tanto nos resulta imposible calcular la media y tendríamos que calcular la mediana.

Aunque la mediana sea siempre la segunda candidata después de la media a la hora de elegir medida de tendencia central, hay ocasiones en que esto no es posible o aconsejable y entonces tengamos que calcular la moda. Este será el caso en que la variable sea nominal y entonces cualquier operación aritmética con los números que representan los valores de la variable está fuera de lugar. El otro caso en que elegiremos la moda se produce cuando la mediana pertenezca a un intervalo abierto. La fórmula para calcular los percentiles supone una distribución uniforme de las puntuaciones en el intervalo y por tanto exige que éste sea cerrado. Si la mediana está dentro de uno de estos intervalos sencillamente no se podrá calcular. Vamos a poner un ejemplo de esta situación: supongamos que queremos medir el tiempo que emplean nuestros alumnos en ejecutar una tarea, es posible que algunos de nuestros alumnos no consigan finalizarla antes de que nos veamos obligados a suspender la sesión y sin embargo, estemos seguros que con más tiempo acabarían la tarea. Entonces un intervalo en la distribución de los tiempos podría ser "Más de una hora", en el caso de que una hora fuese el tiempo máximo disponible para realizar el experimento. Imaginemos que la distribución de frecuencias, expresado el tiempo en minutos fuera la siguiente:

X_i	f_i
Más de 60	25
45-60	10
30-45	5
15-30	3
0-15	1

Tabla 8: Distribución de la variable: “Tiempo empleado en realizar la tarea”

La mitad de los alumnos son 22, entonces la mediana se encontraría en el intervalo "Más de 60", por lo que no podríamos calcularla. Como medida de tendencia central sólo podremos referirnos a la moda y afirmar que lo más frecuente es que los alumnos tarden más de una hora en finalizar la tarea. Si esta tarea hubiese sido un examen tendríamos que reconocer que se trataba de un examen con demasiadas preguntas o demasiado complejas como para realizarlo, en general, en una hora.

Para finalizar diremos que en algunos casos los tres índices de tendencia central son muy parecidos. Si la distribución es simétrica y unimodal los tres coinciden. Cuanto más asimétrica es una distribución más se alejan, la media se desplaza hacia la cola larga de la distribución dejando a la mediana entre ella y la moda, siempre que estemos ante distribuciones unimodales. En cualquier caso, si las tres son muy diferentes se puede hacer referencia a más de uno de ellos para que el lector posea más información.

También podemos poner ejemplos de conjuntos de datos que, simplemente no tienen tendencia central. O, mejor dicho, sus medidas de tendencia central son muy poco representativas del conjunto de datos del cual provienen. Este es el caso de pruebas o test construidos especialmente para discriminar a los alumnos que han adquirido cierta habilidad o algún conocimiento de los que no lo han hecho. Supongamos que esta es la distribución de frecuencias de un test que separa a los alumnos que han aprendido el concepto de límite de una sucesión de los que no:

X_i	f_i
10	20
9	10
8	9
7	8
3	5
2	10
1	20
0	18

Tabla 9: Puntuaciones del test.

La media de estas puntuaciones es 4,91 y sin embargo ningún alumno obtuvo una puntuación de 4,91 y las puntuaciones más cercanas (3 y 7) están aproximadamente a una distancia de 2 unidades de ella. Parece que la media no es un buen descriptor de la tendencia central de esta distribución. La mediana es 2,8, la puntuación que divide la distribución en dos partes iguales. Si calculamos los percentiles correspondientes a las siguientes puntuaciones en la escala, que son 3 y 7, resultan el 50,5 y el 57. Este porcentaje nos hace sospechar de la mediana puesto que puntuaciones lejanas en la escala, 2,8 y 7, resultan muy cercanas en los percentiles, 50 y 57. Tampoco la mediana es una buena representación de los datos. En cuanto a la moda, esta distribución es bimodal, con dos modas absolutas que corresponden al uno y al diez. Y quizás esta afirmación sea lo más sensato que se puede decir de esta distribución, puesto que precisamente el test se ha construido para distinguir a los alumnos que han adquirido un concepto de los que no y por ello no caben términos medios.

La Figura 12 es una representación del polígono de frecuencias de la distribución de puntuaciones del test. En ella podemos ver reflejadas las afirmaciones acerca de la media y la mediana, puesto que, sabiendo tan sólo que la media es 4,91 y la mediana 2,8, difícilmente podemos imaginar una distribución como esta. Mientras que el hecho de ser bimodal nos permitiría tener una idea más aproximada.

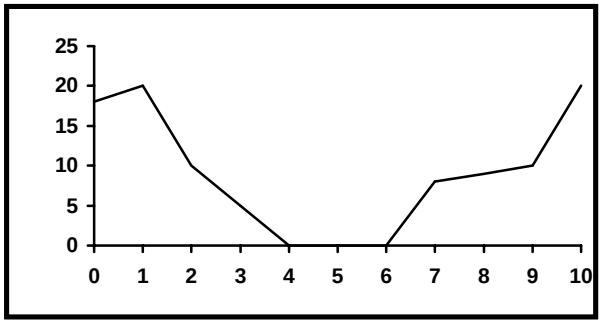

Figura 12: Polígono de frecuencias de la Tabla 9.

Capitulo 4. Medidas de variabilidad

1.- El rango y el rango semiintercuartil

A continuación presentamos dos series, A y B, de diez puntuaciones.

A: 35, 38, 38, 38, 39, 40, 41, 43, 44 y 44

B: 18, 20, 27, 38, 38, 41, 48, 54, 55 y 61.

Proponemos al lector que calcule el valor de la moda, la mediana y la media de cada una de las dos muestras.

La resolución de este pequeño ejercicio deja clara la necesidad de establecer otro tipo de medidas distintas de las de tendencia central para describir una distribución, ya que la muestra A y la muestra B tienen las mismas medias, medianas y modas y, evidentemente, son dos distribuciones distintas. Pero distintas ¿en qué?. Distintas en cuanto a su variabilidad o dispersión, en cuanto al grado en que sus datos se parecen o se diferencian entre sí: mientras que las valoraciones en la muestra A varían entre 35 y 44, en la muestra B lo hacen entre 18 y 61. Por tanto las puntuaciones de esta última se encuentran mucho más dispersas que las de la muestra A.

La forma más sencilla de calcular la variabilidad de un conjunto de datos es hallar la diferencia entre el valor más grande y el valor más pequeño. A este índice se le llama rango de la variable o amplitud total o recorrido. Las valoraciones de la muestra A del problema anterior tienen un rango de 44 - 35 = 9, mientras que el de la muestra B es de 61 - 18 = 43 . Este índice ya nos apunta la diferencia entre las dos distribuciones. Su principal ventaja, la hemos comentado anteriormente, es su facilidad de cálculo y su principal inconveniente es que sólo es sensible a los valores extremos y no se ve afectado en absoluto por los valores centrales. Veamos un ejemplo con tres grupos de puntuaciones:

A: 3, 7, 8, 9, 10, 11, 12, 13	rango = 10
B: 7, 7, 8, 9, 10, 11, 12, 13	rango = 6
C: 7, 10, 10, 10, 10, 10, 10, 13	rango = 6

Las puntuaciones de A y B se parecen en cuanto a dispersión mucho más de lo que se parecen las de B y C, aunque su amplitud total sea la misma. Otro inconveniente del rango es que depende bastante del tamaño de la muestra, si comparamos la dispersión de dos conjuntos de datos de tamaños muy distinto, lo más probable es que la muestra de mayor tamaño tenga también mayor amplitud o recorrido.

Estos dos inconvenientes hacen que el rango no sea una buena medida de variabilidad en solitario, aunque se pueda añadir como complemento a algún otro índice de dispersión.

Otra medida de variabilidad es el llamado rango o amplitud semi-intercuartil, tiene la ventaja sobre el rango de que elimina el influjo de las puntuaciones extremas ya que se calcula mediante los cuartiles primero y tercero y su fórmula es:

$$Q = \frac{Q_3 - Q_1}{2}$$

Su cálculo es más complicado que el del rango pero es más probable que dos distribuciones con el mismo rango semi-intercuartil tengan parecida variación que dos distribuciones con el mismo rango. De hecho, si las distribuciones son simétricas o aproximadamente simétricas el 50% de las observaciones se encuentran entre *Md* - Q y *Md* + Q. Este índice se utiliza sobre todo cuando por las características de los datos utilizamos la mediana como medida de la tendencia central.

2.- Desviación media, desviación estándar y varianza.

En el apartado 5 del capítulo 2, en que describimos las características generales de una distribución de frecuencias, definimos las medidas de variabilidad o dispersión como aquellas que daban cuenta de la concentración o dispersión de los datos en torno a la tendencia central. Así pues, cuando se trata de definir una de estas medidas parece lógico pensar en una expresión que indique de alguna manera un promedio de distancias de las observaciones a la media. Ya vimos que una de las propiedades de la media era que la suma de estas diferencias era cero, es decir:

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0$$

Para obviar este problema tenemos dos posibles soluciones: en primer lugar podemos calcular la suma de los valores absolutos de esas diferencias

$$\sum_{i=1}^{n}\left|X_i - \overline{X}\right|$$

y así, al eliminar los signos, esta suma ya no es igual a cero o bien, en segundo lugar, podemos calcular la suma de los cuadrados de las diferencias

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Aun así, los resultados de estas sumas dependerían del número de elementos que componen la muestra, es decir, estas cantidades son más grandes cuanto mayor sea el tamaño de la muestra (recuérdese que se están sumando números positivos, o mejor, siendo rigurosos, números no negativos). Para evitar este inconveniente, podríamos calcular una "distancia promedio", es decir, dividir cada uno de estos sumatorios por el número de individuos de la muestra.

El primer camino nos lleva a la definición de desviación media, que se calcula, para datos sin agrupar, como:

$$DM = \frac{\sum_{i=1}^{n}\left|X_i - \overline{X}\right|}{n}$$

Y el segundo a la definición de la varianza como:

$$S_X^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

La razón de que el denominador de la varianza sea n-1 y no n, escapa a los conocimientos estadísticos expuestos en estas notas. No obstante podemos dar una sencilla explicación de este hecho: el denominador es n-1 para que la varianza muestral sea una "buena estimación" de la varianza de la población. Se entiende por ser una "buena estimación" aquella que no comete errores sistemáticos.

Vamos a poner un ejemplo para calcular la varianza de dos series de puntuaciones cuya media es 12:

A: 4, 10, 12, 14, 20

B: 10, 11, 12, 13, 14

$$S_A^2 = \frac{\sum_{i=1}^{5}\left(X_i - 12\right)^2}{4} = \frac{(4-12)^2 + (10-12)^2 + (14-12)^2 + (20-12)^2}{4} = \frac{136}{4} = 34$$

$$S_B^2 = \frac{\sum_{i=1}^{5}\left(X_i - 12\right)^2}{4} = \frac{(10-12)^2 + (11-12)^2 + (13-12)^2 + (14-12)^2}{4} = \frac{10}{4} = 2{,}5$$

La única conclusión que podemos extraer de todo esto es que la varianza de A es mucho más grande que la de B, pero en lo que se refiere a cómo de agrupados están los datos respecto a la tendencia central en la muestra A, la magnitud 34 no nos proporciona demasiada información (¿es una varianza muy grande, grande, pequeña?). La cuestión es cómo valorar el grado de dispersión cuantificado mediante la varianza. En realidad, no tiene mucho sentido hablar en términos absolutos de niveles altos o bajos de dispersión sino, más bien, en términos relativos. La varianza sirve, sobre todo, para comparar el grado de dispersión de dos o más conjuntos de valores en una misma variable. Así, comparando varianzas de la misma variable en poblaciones distintas se pueden hacer afirmaciones del tipo: "la población de hombres presenta una mayor variabilidad en su estatura que la población de mujeres, que son más homogéneas en esa característica".

Aun así, por ejemplo, el valor de 34 no parece que tengan relación con las magnitudes de los datos de A, entre 4 y 20, ni con las de la distancia de cada dato a su media, las mayores distancias son de ocho puntos. Esto es así porque, para calcular la varianza hemos elevado cada distancia al cuadrado, estamos "elevando al cuadrado" la unidad de medida de las puntuaciones originales. Por ello, para retomar las unidades originales de esas distancias, se calcula la raíz cuadrada de la varianza, que se denomina desviación típica . Su fórmula es:

$$S_X = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}}$$

Calculando las raíz cuadrada de 34, obtenemos la desviaciones típica de A que es 5,8. Esta magnitud parece guardar más relación con el concepto de separación promedio de los datos respecto a la tendencia central. En la muestra B, la varianza es 2,5 y, por tanto, su desviación típica es 1,58.

De cualquier modo, hay que indicar que tanto la varianza como la desviación típica son siempre cantidades positivas o nulas (este caso ocurrirá cuando todas las observaciones de la variable tengan el mismo valor).

Para calcular la varianza o la desviación típica, no se usa normalmente esta fórmula porque cuando la media es un número decimal, al elevar al cuadrado cada una de las diferencias obtendremos de nuevo un número con el doble de cifras decimales que tendremos que arrastrar o bien redondear lo que aumenta considerablemente el riesgo de cometer errores. Normalmente la fórmula usada para el cálculo de la varianza es:

$$S_X^2 = \frac{\sum_{i=1}^{n} X_i^2 - n \cdot \overline{X}^2}{n-1}$$

y la de la desviación típica

$$S_X = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - n \cdot \overline{X}^2}{n-1}}$$

Estas fórmulas se obtienen sólo con desarrollar el binomio al cuadrado que aparece en la expresión original.

Al igual que en el caso de la media, también podemos escribir expresiones para la varianza y la desviación típica en caso de contar con una distribución de frecuencias sin agrupar o agrupada en intervalos de clase. En el primer caso dichas expresiones son:

$$S_X^2 = \frac{\sum_{i=1}^{k} X_i^2 \cdot f_i - n \cdot \overline{X}^2}{n-1} \text{ y } S_X = \sqrt{\frac{\sum_{i=1}^{k} X_i^2 \cdot f_i - n \cdot \overline{X}^2}{n-1}}$$

dónde f_i es la frecuencia de la puntuación X_i y k es el número de puntuaciones distintas que se han obtenido.

Y si contamos con una distribución de frecuencias agrupadas en intervalos de clase, trabajaremos con el número de intervalos de clase, sus puntos medios y sus respectivas frecuencias, que denotamos en las siguientes expresiones como k, Xm_i y f_i respectivamente:

$$S_X^2 = \frac{\sum_{i=1}^{k} Xm_i^2 \cdot f_i - n \cdot \overline{X}^2}{n-1} \text{ y } S_X = \sqrt{\frac{\sum_{i=1}^{k} Xm_i^2 \cdot f_i - n \cdot \overline{X}^2}{n-1}}$$

Igualmente se cuenta con distintas expresiones para la desviación media en el caso de contar con una distribución de frecuencias sin agrupar o agrupada en intervalos de clase, su determinación se propone como tarea al alumno.

A continuación pondremos ejemplos de la confección de las tablas de distribución para calcular la varianza y la desviación típica con ayuda de las anteriores expresiones. Retomamos para ello los datos de la Tabla 5, relativos a una distribución de frecuencias sin agrupar, a la que añadiremos una columna con los puntos medios de cada intervalo de clase al cuadrado y otra con el producto de éstos por la frecuencia del intervalo. El resultado aparece reflejado en la Tabla 10. De ella extraemos los siguientes datos:

$$\overline{X} = \frac{94}{20} = 4{,}7 \ , \ \sum_{i=1}^{20} X_i^2 \cdot f_i = 524$$

$$S_X^2 = \frac{524 - 20 \cdot (4{,}7)^2}{19} = 4{,}32 \text{ y } S_X = \sqrt{\frac{524 - 20 \cdot (4{,}7)^2}{19}} = 2{,}07$$

X_i	f_i	$X_i \cdot f_i$	X_i^2	$X_i^2 \cdot f_i$
9	1	9	81	81
8	1	8	64	64
7	2	14	49	98
6	2	12	36	72
5	5	25	25	125
4	3	12	16	48
3	3	9	9	27
2	2	4	4	8
1	1	1	1	1
Total	20	94		524

Tabla 10: Tabla de distribución de frecuencias para el cálculo de la desviación típica

Para ejemplificar el cálculo de varianza y desviación típica en el caso de una distribución de frecuencias agrupada en intervalos de clase, volvemos sobre las puntuaciones de 30 alumnos en el test de hábitos de estudio. En concreto, completaremos la Tabla 6 con nuevas columnas: la de los puntos medios de cada intervalo al cuadrado y la de los productos de éstos por la frecuencia del intervalo. El resultado es la Tabla 11.

X_i	Xm_i	f_i	$Xm_i \cdot f_i$	Xm_i^2	$Xm_i^2 \cdot f_i$
90-94	92	2	184	8464	16928
85-89	87	2	174	7569	15138
80-84	82	1	82	6724	6724
75-79	77	4	308	5929	23716
70-74	72	5	360	5184	25920
65-69	67	2	134	4489	8978
60-64	62	3	186	3844	11532
55-59	57	3	171	3249	9747
50-54	52	2	104	2704	5408
45-49	47	2	94	2209	4418
40-44	42	1	42	1764	1764
35-39	37	1	37	1369	1369
30-34	32	2	64	1024	2048
Total		30	1940		133690

Tabla 11: Tabla de distribución de frecuencias para el cálculo de la varianza y la desviación típica

Con los datos de la Tabla 11, ya estamos en condiciones de aplicar las expresiones para el cálculo de la varianza y la desviación típica:

$$\overline{X} = \frac{1940}{30} = 64{,}67\,,\ \sum_{i=1}^{13} Xm_i^2 \cdot f_i = 133690$$

$$S_X^2 = \frac{133690 - 30 \cdot (64{,}67)^2}{29} = 283{,}58 \text{ y } S_x = \sqrt{\frac{133690 - 30 \cdot (64{,}67)^2}{29}} = 16{,}84$$

3.- Uso de las medidas de variabilidad.

Habitualmente cuando nos planteamos el cálculo de una medida de variabilidad para una distribución solemos elegir la desviación típica, sobre todo si se quieren emplear más adelante otras técnicas estadísticas.

Para calcular la desviación típica es necesario conocer la media. Cuando esto no sea posible y sólo podamos calcular la mediana como medida de tendencia central, o si la distribución está truncada o incompleta, usaremos el rango intercuartil como medida de variabilidad.

Cuando estemos ante una distribución con variaciones extremas, podemos calcular la desviación media.

Cuando se quiere comparar la variabilidad de grupos con medias muy diferentes no resulta apropiado comparar sus varianzas sino más bien comparar el que se denomina coeficiente de variación que se representa como CV y se calcula:

$$CV = \frac{S_X}{\overline{X}} \cdot 100$$

Este coeficiente está expresado como un porcentaje y nos da idea de la representatividad de la media. Cuanto mayor es este coeficiente menos representativa es la media. Un ejemplo del uso de este índice se da cuando queremos comparar la variabilidad del tiempo empleado en correr 1500 metros por un grupo de alumnos y otro grupo de alumnas. La diferencia de las medias del tiempo empleado nos aconsejará el uso del coeficiente de variación.

4.- Puntuaciones típicas o estándar

En el capítulo dos hablamos del rango del percentil como un instrumento válido para comparar medidas de una variable en sujetos distintos o medidas de variables distintas en el mismo sujeto. En definitiva el rango del percentil nos sirve para comparar distintas magnitudes tomando como referencia un grupo de individuos a los que se les ha medido esta característica. Ahora volvemos sobre este problema con otra herramienta: las puntuaciones típicas que carecen del inconveniente de ser una escala ordinal como el rango del percentil.

Supongamos que se nos informa que un individuo ha obtenido la puntuación 35 en una prueba objetiva. Si queremos hacer una valoración de ese dato necesitamos conocer algún otro al que podamos hacer referencia. Si conocemos la media de las puntuaciones de los alumnos de la clase, supongamos que la media es 30, entonces podemos calcular la diferencia y decir que la puntuación del alumno en cuestión está 5 puntos por encima de la media. A esta diferencia se la conoce como puntuación diferencial, y dependiendo de su signo, al menos nos informa si la puntuación original está por encima o por debajo de la media. Pero estas puntuaciones diferenciales no son suficientes para comparar. Imaginemos que esta prueba objetiva se ha realizado en los grupos A y B, la media ha sido 30 en los dos mientras que la desviación ha sido 5 en el A y 10 en el B. Tenemos dos alumnos, uno de cada grupo, que han obtenido la puntuación 35. ¿Tiene el mismo valor en el grupo A y en el B?. En principio no, puesto que en el grupo A las puntuaciones son mucho más homogéneas que en el B y, por lo tanto, el valor 35 será una puntuación más extrema, más alta en A que en B.

Una solución a este problema de interpretación consiste en no medir las distancias a la media en términos absolutos, sino con relación al grupo de referencia. Se trataría de indicar cómo de grande es una distancia en términos de las distancias observadas en general en esas puntuaciones. Esa distancia general la habíamos medido en el capítulo anterior mediante la desviación típica y, por tanto, podemos utilizar ésta como unidad de medida. Las puntuaciones así obtenidas se denominan puntuaciones típicas o estándar y se representan por la letra z y su fórmula es:

$$z_i = \frac{X_i - \overline{X}}{S_X}$$

Al proceso de obtención de puntuaciones típicas se llama tipificación. En el caso de nuestros alumnos, el del grupo A tiene una puntuación típica de 35 - 30 / 5 = 1 y el del grupo B de 35 - 30 / 10 = 0,5. Ahora quedan patentes las diferencias entre ambas

puntuaciones: el primero se separa de la media de su grupo en una desviación, mientras que el segundo lo hace en media desviación. Dicho de otra forma, el primer alumno ha destacado bastante más que el segundo dentro de su clase.

Podemos dar como definición de una puntuación típica la siguiente: la puntuación típica de una observación indica el número de desviaciones típicas que esa observación se separa de la media del grupo de observaciones.

Las puntuaciones típicas nos sirven para comparar puntuaciones de individuos distintos en la misma variable y también puntuaciones de distintas variables en individuos distintos siempre calculadas respecto a los parámetros de su grupo de referencia.

La tipificación resulta además de gran utilidad puesto que la media y la desviación típica de una variable tipificada son siempre 0 y 1 respectivamente, sean cuales sean estos parámetros en las variables de las que proceden. La demostración es muy sencilla:

$$\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{X_i - \overline{X}}{S_X} = \frac{1}{n \cdot S_X} \cdot \sum_{i=1}^{n} X_i - \frac{1}{n \cdot S_X} \cdot \sum_{i=1}^{n} \overline{X} = \frac{1}{n \cdot S_X} \cdot n \cdot \overline{X} - \frac{1}{n \cdot S_X} \cdot n \cdot \overline{X} = 0$$

$$S_z^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (z_i - \bar{z})^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{S_X^2} = \frac{1}{(n-1) \cdot S_X^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{(n-1) \cdot S_X^2} \cdot (n-1) \cdot S_X^2 = 1$$

Las puntuaciones típicas reflejan, en cierto sentido, las relaciones esenciales entre las puntuaciones, con independencia de la unidad de medida utilizada en la medición: algo así como la estructura atómica de esas relaciones.

5.- Escalas derivadas.

Hemos mencionado anteriormente las ventajas que tienen las puntuaciones típicas pero también tienen algún inconveniente: el más importante es que al tener media cero y desviación típica de uno, muchas de las puntuaciones típicas que calculemos serán números negativos y decimales. Un procedimiento para evitar el manejo de estas cantidades consiste en transformar linealmente las puntuaciones típicas de forma que sean puntuaciones equivalentes a las originales y que constituyen lo que se denomina una escala derivada.

Supongamos que tenemos un conjunto de n puntuaciones originales X_i, a partir de estas calculamos las puntuaciones típicas z_i, y después realizamos una transformación lineal del tipo $H_i = a \cdot z_i + b$ donde a y b son dos constantes. Las puntuaciones H_i así obtenidas son una escala derivada equivalente a la de las puntuaciones típicas. La media de estas puntuaciones será b ,la desviación típica |a| y la varianza a^2.

Alguna de estas escalas tienen nombre propio por su amplio uso sobre todo en el campo de la Psicología: por ejemplo las puntuaciones T tienen media 50 y desviación

típica 10, es decir se calculan $T_i = 10 \cdot z_i + 50$, o la escala de los estaninos que tiene media 5 y desviación típica de 2, $S_i = 2 \cdot z_i + 5$.

Capitulo 5. Correlación lineal y regresión lineal simple

1.- Introducción

Uno de los objetivos principales de la ciencia consiste en descubrir las relaciones entre variables y la estadística cuenta con herramientas eficaces para llevar a cabo esta tarea. La observación de relaciones entre variables ayuda a comprender los fenómenos y a encontrar explicaciones de los mismos, e indica las vías probablemente más eficaces para intervenir sobre las situaciones. Así, por ejemplo, un profesor se puede preguntar si los alumnos de clases más numerosas adquieren menos conocimientos que los de las menos numerosas, si los alumnos con buenos conocimientos previos obtienen, en general, mejores calificaciones que los que no los poseen o qué tipo de presentaciones de contenidos están asociados con un mantenimiento continuado de la atención.

Desde el punto de vista matemático, las relaciones entre variables pueden ser de muchos tipos, por ejemplo las siguientes expresiones indican relaciones entre variables:

$$Y= 3X + 2 \qquad Y= 8^X \qquad Y= 1/X \qquad Y=X^3$$

Estas funciones como conceptos matemáticos son teóricas e ideales. Cuando recogemos datos de nuestros alumnos no nos encontramos con relaciones de este estilo sino con conjuntos de datos que muestran una configuración concreta, y lo que nos preguntamos es si esa configuración se parece a alguno de esos modelos teóricos o, lo que es lo mismo, si ese modelo explica bien la relación entre esas variables.

En estas notas nos limitaremos a estudiar relaciones lineales entre variables, es decir, relaciones que se ajusten al modelo $Y = a X + b$, dónde a y b son constantes. Se llaman lineales porque la representación gráfica de Y es una recta. De forma que a partir de este párrafo, cuando hablamos de relaciones, en realidad estamos hablando sólo de relaciones lineales.

Una vez conocidos el grado y el sentido de una relación, podríamos ir un poco más allá. Por ejemplo, si encontramos relación entre inteligencia y rendimiento, podríamos, en el caso de que para un sujeto conociésemos sólo su puntuación en el test de inteligencia, hacernos una idea de su rendimiento al final del curso. Por el contrario, y dada la ausencia de relación entre inteligencia y estatura, de nada nos serviría conocer la estatura de una persona para predecir su inteligencia.

En definitiva, el otro problema que vamos a abordar en este capítulo es la utilización de la información que nos aportan las relaciones lineales observadas entre variables para, conociendo la puntuación de un sujeto en una de ellas, pronosticar su puntuación en la otra. Así, el término regresión se utiliza como sinónimo de predicción por razones de tipo histórico, ya que fue Francis Galton quien desarrolló estas técnicas para el estudio de la herencia de algunos caracteres como la estatura. Una de las conclusiones de estos estudios fue que la estatura de los hijos respecto de la de sus padres sufre una regresión a la media, es decir, los hijos de padres con una determinada altura tienen una estatura media más cercana a la media de la población que a la de sus padres. De hecho, también la notación como r del coeficiente de correlación de Pearson tiene este mismo origen.

2.- Tipos de relaciones lineales

Dos variables, X e Y, tienen una relación lineal directa cuando los valores altos en Y tienden a emparejarse con valores altos en X, los valores intermedios de Y tienden a emparejarse con valores intermedios en X, y los valores bajos de Y tienden a emparejarse con niveles bajos en X. Un ejemplo: este tipo de relación es la que se da entre la variable X (puntuación en un test de inteligencia) y la variable Y (nota media de las calificaciones de las asignaturas al final de curso).

Dos variables, X e Y, mantienen una relación lineal inversa cuando los valores altos en Y tienden a emparejarse con valores bajos en X, los valores intermedios en Y tienden a emparejarse con valores intermedios en X y los valores bajos en Y tienden a emparejarse con valores altos en X. Si medimos en nuestros alumnos el tiempo invertido en realizar una tarea, variable X, y el número de errores cometidos al realizarla, variable Y, seguramente observaremos que entre estas dos variables se da una relación lineal inversa.

Se dice que no hay relación lineal ó que la relación lineal es nula entre dos variables cuando no hay un emparejamiento sistemático entre ellas en función de sus valores. Si tenemos datos de las puntuaciones obtenidas por nuestros alumnos en un test de ansiedad y de sus tallas, observaremos que entre estas dos variables no hay relación lineal.

3.- Representación gráfica de relaciones entre variables

Una forma sencilla para averiguar si entre dos variables, X e Y, existe relación y, en su caso, determinar el tipo de la misma, consiste en la realización de un gráfico en el que aparecen las dos variables de una forma conjunta. Como en todos los gráficos, en primer lugar determinaremos los dos ejes de coordenadas de forma que, en cada uno de ellos, se represente una de las dos variables. A continuación, para cada par de valores (x,y), se dibuja un punto con las correspondientes coordenadas, es decir, a una distancia x del eje de abcisas y una distancia y del eje de ordenadas. El resultado es un gráfico llamado diagrama de dispersión, diagrama de puntos o nube de puntos.

Cuando entre dos variables hay una relación lineal directa, la nube de puntos suele tener una forma alargada (tendencia a la linealidad) e inclinada hacia arriba; puesto que valores altos de la variable X se emparejan con valores altos de la variable Y, valores bajos de X se emparejan con valores bajos de Y y valores intermedios con valores intermedios.

Cuando la relación lineal es inversa también se observa una tendencia a la linea-lidad pero inclinada hacia abajo. Esto es así porque los valores altos de X se emparejan con valores bajos de Y, los valores bajos de X se emparejan con valores altos de Y y valores intermedios se emparejan con valores intermedios.

Si no hay relación lineal los puntos pueden aparecer dispersos en el espacio o mostrar otro tipo de tendencia distinta a la linealidad, por ejemplo pueden tender a situarse en una parábola. Recordemos que lo que sucede en este caso es una ausencia de emparejamiento sistemático entre los valores de las variables.

En la Tabla 12 aparecen los datos de las seis variables citadas anteriormente medidas en 15 alumnos. Los gráficos que siguen son los diagramas de dispersión de los tres pares de variables. En la Figura 15 se aprecia que la nube de puntos tiene una forma alargada e inclinada hacia arriba, la relación entre "Inteligencia" y "Rendimiento", es lineal directa. En la Figura 16, por el contrario, aparece una relación lineal inversa entre "Tiempo invertido en realizar una tarea" y "Número de errores cometidos". Por último, como sospechábamos, en la Figura 17 no podemos apreciar ninguna tendencia en la nube de puntos que representan las puntuaciones en "Ansiedad" y la "Estatura"

Sujeto	Inteli.	Rendto.	Tiempo	Errores	Estatura	Ansiedad
1	9	5	7	4	7	3
2	12	5	11	2	8	1
3	6	1	5	4	5	3
4	9	4	5	5	12	3
5	7	2	6	4	8	2
6	9	2	9	4	9	4
7	5	1	13	1	7	4
8	9	3	8	2	6	4
9	7	3	4	5	6	3
10	3	1	9	3	9	2
11	10	4	6	3	9	3
12	6	2	10	2	6	2
13	11	5	11	1	10	2
14	4	2	9	2	10	4
15	13	5	7	3	8	5

Tabla 12: Puntuaciones obtenidas por 15 alumnos en 6 pruebas distintas.

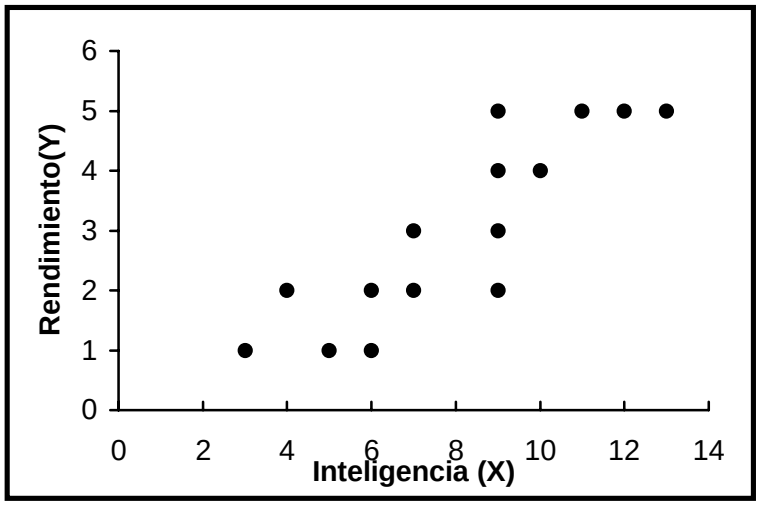

Figura 15: Diagrama de dispersión de las puntuaciones de Inteligencia y Rendimiento de la Tabla 12

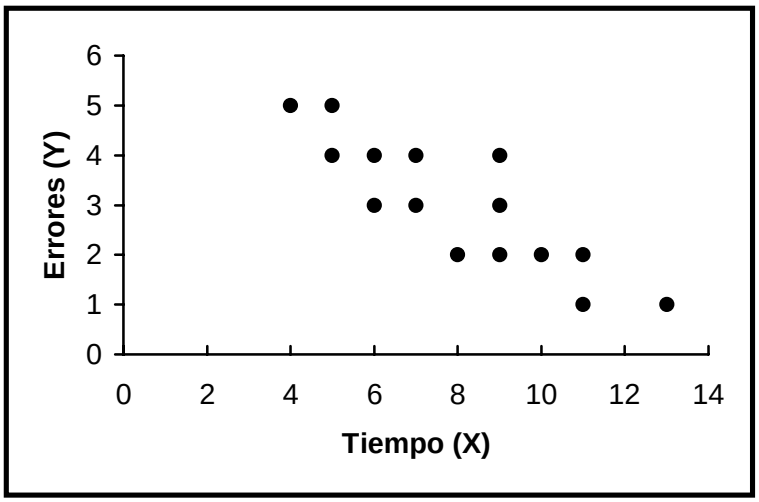

Figura 16: Diagrama de dispersión de las variables "Tiempo empleado en realizar una tarea" y "Número de errores cometidos" reflejadas en la Tabla 12.

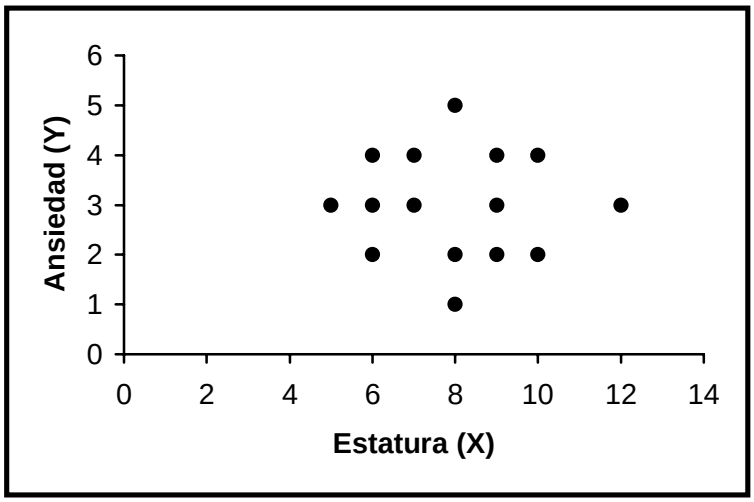

Figura 17: Diagrama de dispersión de las variables "Estatura" y "Ansiedad" medidas en los sujetos que aparecen en la Tabla 12.

4.- Cuantificación de una relación lineal.

Una vez definidos estos tres tipos de relación lineal nos enfrentamos al problema de construir un índice que nos revele el sentido y la magnitud de las mismas. Nos vamos a detener en primer lugar en el promedio de productos cruzados de cada puntuación en X e Y menos sus respectivas medias. La expresión es la siguiente:

$$\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\cdot\left(Y_i-\overline{Y}\right)}{n-1}$$

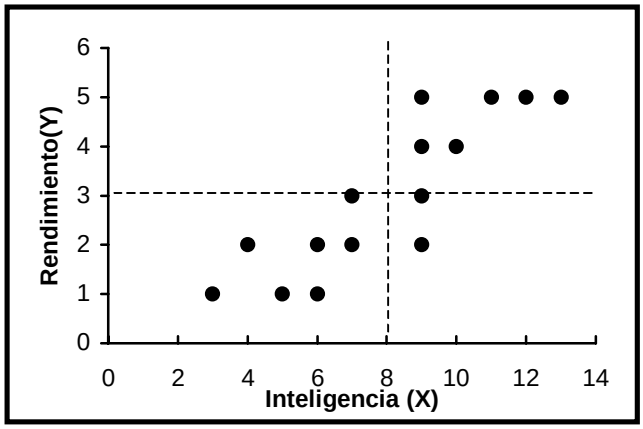

Figura 18: Diagrama de dispersión al que se han añadido dos líneas discontinuas a la altura de las medias de cada una de las dos variables que son 8 para la variable X y 3 para la variable Y.

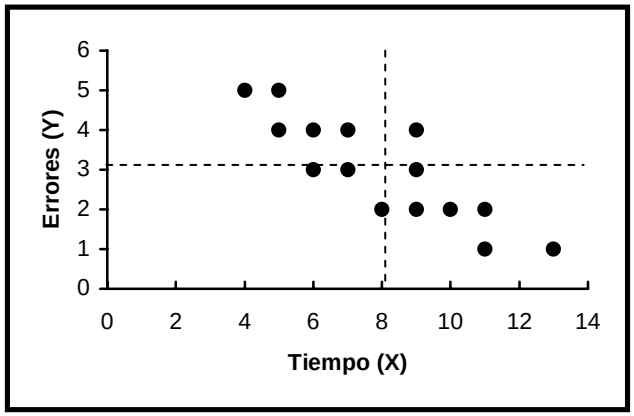

Figura 19: Diagrama de dispersión al que se han añadido dos líneas discontinuas a la altura de las medias de cada una de las dos variables que son 8 para la variable X y 3 para la variable Y.

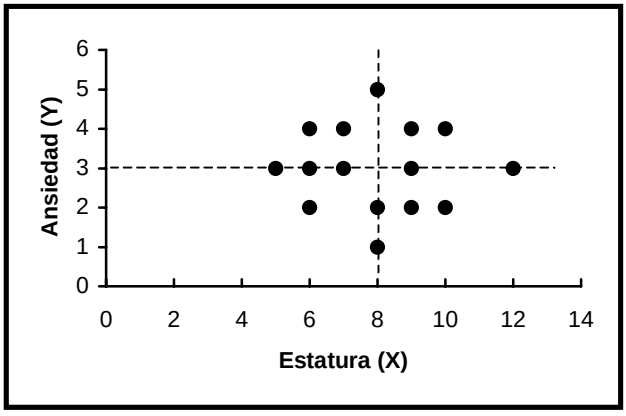

Figura 20: Diagrama de dispersión al que se han añadido dos líneas discontinuas a la altura de las medias de cada una de las dos variables que son 8 para la variable X y 3 para la variable Y.

Con ayuda de los gráficos anteriores vamos a ver cómo este índice puede servir a nuestros propósitos. En cada uno de los diagramas de puntos se han representado mediante dos rectas en trazos discontinuos las medias de cada variable representada de forma que cada gráfico queda dividido en cuatro partes. Hay que señalar que estas medias son 8 para las tres variables representadas como X y 3 para las representadas como Y. Si nos fijamos en la Figura 18, la mayoría de los puntos se sitúan en el cuadrante superior derecho y en el inferior izquierdo. En estos dos cuadrantes los factores que aparecen en la expresión son ambos positivos o negativos (las puntuaciones están tanto en X como en Y por encima de la media o por debajo) por lo que su producto resulta siempre positivo. Como sumamos estos productos para cada par de observaciones, el resultado de esa expresión será grande y positivo.

Si prestamos ahora atención a la Figura 19 vemos que la mayoría de los puntos se encuentran en el cuadrante superior izquierdo y en el inferior derecho. En estos cuadrantes una de las puntuaciones es mayor que la media y la otra inferior, por lo que en la expresión uno de los factores será positivo y otro negativo y en consecuencia su producto será un número negativo. Como vamos sumando estos productos para cada par de observaciones, el resultado será una cantidad grande y negativa.

Los puntos de la Figura 20 se distribuyen casi en igual número por los cuatro cuadrantes por lo que en la expresión habrá prácticamente en igual número productos positivos y negativos. Al efectuar la suma, se compensaran unos con otros y el resultado será una cantidad próxima a cero.

Así, la anterior expresión es positiva si la relación lineal es directa, negativa cuando es inversa y cercana a cero en ausencia de relación. Además, su valor absoluto será más grande cuanto más acusada sea la tendencia a la linealidad en el diagrama de dispersión. Este índice se llama covarianza y se representa por S_{XY}.

Si calculamos las covarianzas para los tres pares de variables anteriores los resultados son: S_{XY} = 3,667, S_{XY} = -2,733 y S_{XY} = -0,067. Como se esperaba, para el primer par la covarianza es positiva, negativa en el segundo caso y cercana a cero en el tercero. Pero aún encontramos un inconveniente, no podemos decir cómo es en cuanto a su magnitud, no podemos decir que 3,667 es un valor grande de covarianza y -0,067 sea insignificante puesto que la covarianza no tiene un máximo y un mínimo que sean comunes a todos los casos. Incluso si cambiásemos de unidad de medida de alguna de las variables la covarianza se vería afectada, mientras que la relación lineal entre las dos variables seguiría siendo la misma.

Una forma de resolver este problema es calcular la covarianza sobre puntuaciones cuya varianza permanezca invariable a cambios en las unidades de medida, es decir, calcularla sobre las puntuaciones típicas o estándar. Este índice fue desarrollado por Francis Galton y Karl Pearson aunque se le conoce como coeficiente de correlación de Pearson y se representa por la letra r ya que se comenzó a utilizar en el estudio de la asociación de características físicas humanas, estudio que por primera vez mostró la naturaleza regresiva de las medidas físicas entre una generación y la siguiente. El coeficiente de correlación de Pearson entre las variables X e Y viene dado por la fórmula:

$$r_{XY} = \frac{\sum z_{Xi} \cdot z_{Yi}}{n} = \frac{S_{XY}}{S_X \cdot S_Y}$$

Esta fórmula no resulta muy práctica a la hora de calcular el coeficiente puesto que habría que tipificar cada una de las puntuaciones en ambas variables, después los productos cruzados y por último la suma de todos esos productos. Para evitar el problema de la tipificación e incluso el del cálculo de la desviación típica se ha desarrollado la siguiente expresión:

$$r_{XY} = \frac{n \cdot \sum_{i=1}^{n} X_i \cdot Y_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{\sqrt{n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}}$$

Vamos a calcular ahora el coeficiente de correlación de Pearson para el primer par de variables como ejemplo del uso de esta última fórmula. Para ello se confecciona una tabla como la siguiente:

X	Y	X^2	Y^2	XY
9	5	81	25	45
12	5	144	25	60
6	1	36	1	6
9	4	81	16	36
7	2	49	4	14
9	2	81	4	18
5	1	25	1	5
9	3	81	9	27
7	3	49	9	21
3	1	9	1	3
10	4	100	16	40
6	2	36	4	12
11	5	121	25	55
4	2	16	4	8
13	5	169	25	65
120	45	1078	169	415

Tabla 13: Puntuaciones de 15 sujetos en las variables X e Y, sus respectivos cuadrados y el producto cruzado.

Ahora sustituimos en la expresión las correspondientes cantidades, teniendo en cuenta que n es 15, tenemos 15 pares de puntuaciones, y los resultados de las sumas aparecen en la última línea de la Tabla 13.

$$r_{XY} = \frac{15 \cdot 415 - 120 \cdot 45}{\sqrt{15 \cdot 1078 - 120^2} \cdot \sqrt{15 \cdot 169 - 45^2}} = 0{,}868$$

Por último, si nos fijamos en la fórmula que define el coeficiente de correlación de Pearson como cociente de la covarianza y el producto de las desviaciones típicas, advertiremos que, puesto que éstas últimas son siempre positivas, el signo del coeficiente es el mismo que el de la covarianza. Y además, si alguna de las varianzas es cero, no podremos calcular el coeficiente de correlación de Pearson, la correlación es inde-terminada.

5.- Propiedades del coeficiente de correlación de Pearson

Nos hemos referido a la dificultad de la covarianza para expresar la magnitud de una relación lineal por carecer de un máximo y un mínimo estables. Pues bien, esta la primera y más importante propiedad a destacar del coeficiente de correlación de Pearson: r siempre toma valores entre -1 y 1. Así, ahora podemos afirmar que el coeficiente entre inteligencia y rendimiento, que tomaba un valor de 0.868 es alto, puesto que el valor mayor que podría tomar es 1.

Una segunda propiedad del coeficiente de correlación es que si hacemos transformaciones lineales en las variables, del tipo $U = a\ X + b$ y $V = c\ Y + d$ con a y c constantes positivas, entonces el coeficiente de correlación no cambia. Es decir, $r_{XY} = r_{UV}$.

Esta propiedad también se cumple cuando ambas constantes son positivas y si una es positiva y otra negativa los coeficientes de correlación son iguales en valor absoluto y cambian en cuanto a su signo.

6.- Interpretación de una correlación.

A la hora de interpretar un coeficiente de correlación hay que tener en cuenta por un lado su magnitud y por otro su signo. La magnitud se refiere al grado en que la relación entre las dos variables queda bien descrita con r, mientras que el signo se refiere al tipo de relación.

Un coeficiente de correlación positivo entre las variables X e Y, indica la tendencia a aumentar los valores de Y cuando aumentamos los de X, y a disminuir los valores de Y cuando disminuimos los de X.

Un coeficiente de correlación negativo indica tendencia a disminuir los valores de Y cuando aumentamos los de X y a aumentar los de Y cuando disminuimos los de X.

Un coeficiente de correlación en torno a cero indica que el modelo de relación lineal entre esas variables no es válido. Que cuando aumentamos X, Y puede indistintamente aumentar o disminuir. Es un error interpretar que un coeficiente de correlación cercano a cero indica independencia de las variables, puesto que este coeficiente sólo expresa relación lineal y ya hemos dicho a principio del capítulo que entre dos variables puede haber otros muchos tipos de relación.

En cuanto a la magnitud del coeficiente de correlación, tradicionalmente se han establecido unos puntos de corte generales de la siguiente forma: si el módulo del

coeficiente de correlación se sitúa entre 0 y 0,20, entonces es insignificante, si está entre 0,20 y 0,50 medio, entre 0,50 y 0,80 alto y a partir de 0,80 muy alto.

La tendencia actual va más en la línea de establecer que en cada área de estudio se va desarrollando un conocimiento que permite valorar los coeficientes de correlación en términos relativos. Por ejemplo, cuando se quiere medir la estabilidad de un test se aplica dos veces el mismo test en un intervalo corto de tiempo a los mismos sujetos. Si el test es estable, los sujetos contestarán aproximadamente igual en las dos ocasiones. Por tanto, si hallamos el coeficiente de correlación entre las puntuaciones obtenidas en las dos administraciones del test, habremos calculado un índice de fiabilidad: cuanto mayor sea el coeficiente más fiable será el test y viceversa. Pues bien, este índice de fiabilidad debe ser al menos de 0,80. Es decir, en este caso el punto de corte es 0,80, si el coeficiente es menor el test no es estable. Mientras que para investigaciones relacionadas con la personalidad, una correlación en torno a 0,30 se puede considerar como muy importante.

Otro error que se suele cometer con bastante frecuencia es el de la interpretación del coeficiente de correlación como un coeficiente de causación, es decir, interpretar que cuando entre dos variables se da una correlación alta, entonces una de ellas es causa de la otra. Cuando en realidad esto puede ocurrir o no, puede haber otras variables no incluidas en el estudio que sean la causa de esas dos. Es decir el coeficiente de correlación expresa relación lineal pero no causalidad entre las variables. Por ejemplo, si medimos en los paises de la ONU la correlación entre el número de coches por cada mil habitantes y el nivel cultural medio de los mismos probablemente resultará una correlación alta y positiva. Esto no quiere decir que la posesión de un coche incremente el nivel cultural de su propietario, ni mucho menos que si en un país cualquiera regalásemos coches a sus habitantes su nivel cultural medio ascendería. Lo que ocurre realmente es que hay al menos una tercera variable que puede ser el nivel de desarrollo económico y social del país que tiene efectos sobre las otras dos variables y que probablemente es realmente su causa.

7.- Una situación prototipo del uso de las técnicas de regresión

Una vez que hemos estudiado la existencia de relación entre dos variables X e Y, estamos en condiciones de plantearnos si sería posible predecir los valores que obtendrán los sujetos en la variable Y cuando conocemos los que han obtenido en la variable X, o viceversa. Este problema se resuelve mediante las técnicas de regresión, entendido este último término como sinónimo de predicción.

Las técnicas de regresión se aplican cuando tenemos observaciones de dos variables X e Y en un conjunto de sujetos y más tarde se incorpora un nuevo sujeto al conjunto del que sólo podemos conocer su puntuación en una variable, por ejemplo X, y queremos hacer un pronóstico de cuál será su puntuación en Y.

Para hacer el pronóstico tenemos varios procedimientos:

1º) Calcular la media de la variable Y y asignar al nuevo sujeto el valor de la media. En este caso no tenemos en cuenta la relación entre las variables ni la puntuación en X del individuo. De hecho, asignaríamos a todos los sujetos que se incorporasen al conjunto la misma puntuación.

2º) Buscar en el conjunto de observaciones si hay algún sujeto cuya puntuación en X sea la misma o cercana a la del nuevo sujeto, entonces pronosticaremos para éste un valor de Y igual al del individuo conocido. Actuando de este modo no explotamos la relación entre las variables X e Y, puesto que el pronóstico se basa sólo en un par de observaciones y no en todo el conjunto.

Ninguno de estos inconvenientes presentan las técnicas de regresión porque se sirven de la información del conjunto de pares de valores y de la relación entre X e Y. De hecho, este tipo de predicción se llama regresión simple, porque nos basamos en una única variable predictora. Si utilizásemos más de una variable predictora hablaríamos de técnicas de regresión múltiple. Además, también vamos a restringir el tipo de relaciones estudiadas entre las variables a las lineales porque son las más sencillas. Así que, más concretamente, vamos a hablar de técnicas de regresión lineal simple.

8.- Regresión simple

En primer lugar vamos a definir algunos términos. Llamamos variable predictora a la que se utiliza para hacer pronósticos y variable criterio a aquella en la que se hacen los pronósticos. Lo que vamos a determinar es la recta de regresión de Y sobre X, es decir, aquella que permite predecir los valores de Y a partir de los de X. Para ello comenzamos por plantearnos la ecuación de una recta $Y' = A_{YX} + B_{YX} X$. Hemos puesto subíndices a las constantes A y B para indicar que estamos determinando la recta de regresión de Y sobre X. Si nos planteáramos determinar la recta de regresión de X sobre Y, estaríamos ante una ecuación del tipo $X' = A_{XY} + B_{XY} Y$. El lector puede imaginar que el procedimiento para obtener las constantes es el mismo aunque, lógicamente, cambie el resultado, es decir, A_{XY} , B_{XY} y A_{YX} , B_{YX} no son las mismas.

El criterio para determinar estas constantes es el de los mínimos cuadrados que consiste en buscar A_{YX} y B_{YX} de modo que minimicen la siguiente expresión:

$$\frac{\sum_{i=1}^{n}(Y_i - Y_i')^2}{n}$$

dónde $Y_i' = A_{YX} + B_{YX} \cdot X_i$

Dicho de otra forma, las constantes A_{YX} y B_{YX} se determinan de modo que el promedio de los errores al cuadrado que cometemos estimando los valores de la variable Y a través de la recta de regresión sea mínimo. Si calculásemos el promedio de errores sin elevarlos al cuadrado, estaríamos sumando cantidades positivas y negativas con lo que, al compensarse, obtendríamos una magnitud engañosa del error. Elevando cada término del error al cuadrado estamos sumando siempre cantidades positivas o, lo que es lo mismo, tenemos en cuenta el tamaño del error y no si éste es por defecto o por exceso. Para que la primera expresión sea mínima se debe derivar parcialmente respecto a A_{YX} y B_{YX} e igualar a cero ambas expresiones. Entonces estaremos ante un sistema de dos ecuaciones con dos incógnitas, que una vez despejadas resultan:

$$B_{YX} = \frac{n \cdot \sum_{i=1}^{n} X_i \cdot Y_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2}$$

$$A_{YX} = \overline{Y} - B_{YX} \cdot \overline{X}$$

El lector ya puede imaginar cómo será la expresión de las constantes de la recta de regresión de X sobre Y:

$$B_{XY} = \frac{n \cdot \sum_{i=1}^{n} X_i \cdot Y_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{n \cdot \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i \right)^2}$$

$$A_{XY} = \overline{X} - B_{XY} \cdot \overline{Y}$$

Hemos hablado de la estrecha relación entre correlación y regresión, pues bien, este hecho se refleja también en las expresiones de los coeficientes de las rectas de regresión de la siguiente manera:

$$B_{YX} = r_{XY} \cdot \frac{S_Y}{S_X}$$

$$B_{XY} = r_{XY} \cdot \frac{S_X}{S_Y}$$

Sustituyendo ahora estos valores en las ecuaciones de las rectas de regresión:

$$Y' = r_{XY} \cdot \frac{S_Y}{S_X} \cdot \left(X - \overline{X} \right) + \overline{Y}$$

$$X' = r_{XY} \cdot \frac{S_X}{S_Y} \cdot \left(Y - \overline{Y} \right) + \overline{X}$$

Si el coeficiente de correlación es 1 ó -1 las dos rectas coinciden, puesto que entonces las observaciones tienen una relación lineal perfecta, están sobre una línea recta que naturalmente es la de regresión.

Por el contrario, si r = 0 entonces las dos rectas de regresión son perpendiculares entre sí y paralelas a cada uno de los ejes de coordenadas puesto que sus ecuaciones serían:

$$Y' = \overline{Y}$$

$$X' = \overline{X}$$

Las dos rectas forman un ángulo agudo cuando el módulo de r es distinto de cero y estrictamente menor que 1.

La relación entre correlación y predicción queda patente cuando se define el coeficiente de determinación, que es igual al coeficiente de correlación elevado al cuadrado, e indica la proporción de varianza del criterio que queda explicada por ese modelo lineal. Es decir, r^2 nos da idea de si el modelo lineal elegido es apropiado o no a nuestros datos.

9.- Algunos índices del error de estimación

Naturalmente, salvo en el caso de que el módulo de r sea 1 (si r = 1 o r = -1), cuando calculamos los valores de la variable criterio usando la recta de regresión estamos cometiendo errores. Cada vez que calculamos un valor Y_i' a través de la recta de regresión, cometemos un error que será igual a la diferencia entre el valor real Y_i y el estimado Y_i'. Precisamente, la varianza de la variable Y-Y' es lo que llamamos error cuadrático medio que notaremos como S_{YX}^2 (en el caso de trabajar con la recta de regresión de X sobre Y será S_{XY}^2 y su expresión análoga) y se calcula como:

$$S_{YX}^2 = \frac{\sum_{i=1}^{n} (Y_i - Y_i')^2}{n-1}$$

Desarrollando el binomio resulta:

$$S_{YX}^2 = S_Y^2 \cdot \left(1 - r_{XY}^2\right)$$

La raíz cuadrada de esta expresión se define como el <u>error estándar de estimación o predicción</u>, se denota S_{YX} y se calcula:

$$S_{YX} = S_Y \sqrt{1 - r_{XY}^2}$$

Para finalizar, haremos hincapié en el hecho de que estos índices son medidas promedio del error cometido cuando estimamos una variable a partir de otra, no indican que siempre cometamos ese error, en algunos casos será mayor y en otros menor.

10.- Un ejemplo de aplicación del modelo de regresión simple

Volvemos ahora sobre un ejemplo ya citado anteriormente, se trata de las variables X: Puntuación en un test de inteligencia e Y: Nota media obtenida al final de curso. La Tabla 13 nos ha servido en el apartado 4 de este mismo capítulo para calcular el coeficiente de correlación de Pearson que resultaba 0,868. Además sabemos que la media de X es 8 y la de Y es 3.

El coeficiente de determinación es el de correlación al cuadrado que resulta 0,75, es decir, el 75% de la varianza de la variable criterio queda explicada por la variable predictora, lo cual nos lleva a pensar que estamos ante unas condiciones apropiadas para aplicar técnicas de regresión lineal.

Ahora vamos a calcular las ecuaciones de la recta de regresión de Y sobre X. Para ello aplicamos las ecuaciones que determinan sus coeficientes con los datos de la Tabla 13:

$$B_{YX} = \frac{15 \cdot 415 - 120 \cdot 45}{15 \cdot 1078 - 120^2} = 0{,}46$$

$$A_{YX} = 3 - 0{,}46 \cdot 8 = -0{,}7$$

Por tanto la recta de regresión de Y sobre X tiene la siguiente ecuación:

$$Y' = 0{,}46 \cdot X - 0{,}7$$

Imaginemos ahora que queremos saber qué nota media hubiera obtenido un alumno que tuvo una puntuación de 8 en el test de inteligencia. Para ello no hay más que sustituir X por 8 en la ecuación, y así el valor de Y' resulta 0,46·8 - 0.7 = 2,98. Vamos a ver que error cometeríamos si para el alumno que obtuvo un 10 en el test de inteligencia calculásemos su rendimiento medio a través de la recta de regresión, es decir, sustituimos en esa ecuación X por 10 y el resultado es 0,46·10-0,7 = 3,9. El valor real es 4, con lo que el error cometido es pequeño, 4-3,9 = 0,1.

Para calcular ahora el error estándar de estimación necesitamos conocer el valor de la desviación típica de Y, que igualmente obtendremos de la Tabla 13:

$$S_Y = \sqrt{\frac{169 - 15 \cdot 9}{14}} = 1{,}56$$

$$S_{YX} = 1{,}56 \cdot \sqrt{1 - 0{,}868^2} = 0{,}77$$

Una estimación del error medio que se comete al predecir Y a partir de X es 0,77, ya hemos visto que el error cometido para X = 10 es 0,1. Para X = 9, el valor de Y resultante de la ecuación de la recta de regresión es 3,44. El primer sujeto de la Tabla 13 obtiene una puntuación de 9 en X y de 5 en Y, por lo que, en este caso, el error cometido sería 5 - 3,44 = 1,56. Así, queda patente que son dos conceptos distintos: 0,77 indica error medio mientras que 0,1 ó 1,56 son errores que cometemos al predecir un sólo dato.

Ahora estableceremos la recta de regresión de X sobre Y, para ello calculamos sus coeficientes:

$$B_{XY} = \frac{15 \cdot 415 - 120 \cdot 45}{15 \cdot 169 - 45^2} = 1{,}62$$

$$A_{XY} = 8 - 1{,}62 \cdot 3 = 3{,}14$$

La ecuación de la recta de regresión de Y sobre X es la siguiente:

$$X' = 1{,}62 \cdot Y + 3{,}14$$

Supongamos ahora que sabemos que la nota media de un alumno ha sido 4 y no hizo en su día el test de inteligencia. Nos gustaría saber qué puntuación hubiera obtenido en ese test, para ello utilizamos la recta de regresión de X sobre Y, sustituimos Y por 4 en la ecuación: 1,62·4 + 3,14 = 9,62. La puntuación esperada en el test de inteligencia es 9,62. Si nos fijamos en la tabla hay dos alumnos que tienen también un 4 como nota media y sus puntuaciones en el test de inteligencia fueron 9 y 10 respectivamente.

El error estándar de estimación vendrá dado por la siguiente ecuación:

$$S_{XY} = S_X \cdot \sqrt{1 - r_{XY}^2}$$

Necesitamos conocer antes la desviación típica de X, a partir de la Tabla 13:

$$S_X = \sqrt{\frac{1078 - 15 \cdot 64}{14}} = 2{,}9$$

$$S_{XY} = 2{,}9 \cdot \sqrt{1 - 0{,}868^2} = 1{,}44$$

En este caso, el error estándar de estimación es mayor que en el caso de estimar Y a partir de X, la razón es muy sencilla, este error depende de la desviación típica de cada variable y del coeficiente de correlación. Cuanto mayor es la desviación típica de la variable criterio mayor será el error, dicho de otra forma, se hacen peores predicciones de variables dispersas que de variables homogéneas. En el ejemplo, la desviación típica de X es mayor que la de Y por tanto el error estándar de estimación es mayor que cuando estimamos Y a partir de X.

Cuanto mayor es el módulo del coeficiente de correlación más pequeño será el error estándar de estimación. Podríamos decir que cuanto más fuerte es la relación lineal entre variables, mejores serán las predicciones basadas en ellas.

ÍNDICE