Comentarios de

Estadística Descriptiva

con una y dos variables

David Casado

Universidad Complutense de Madrid
∟ Facultad de Ciencias Económicas y Empresariales
∟ Departamento de Estadística e Investigación Operativa II
∟ David Casado de Lucas

15 febrero 2012

Puedes no imprimir este archivo y consultarlo en formato digital, ahorrarás papel y tinta. Si decides imprimirlo, por favor hazlo en papel reciclado, a doble cara y con poca tinta. Sé ecológico. Muchas gracias.

Prólogo

En este archivo se incluyen algunos comentarios relacionados con la Estadística descriptiva. El origen de estos comentarios está en mi docencia, como doctorando, en la Universidad Carlos III de Madrid. La mayor parte de estas notas ha sido escrita para los alumnos de la asignaturas de Estadística de la Licenciatura en Documentación y del Grado en Información y Documentación, ambos planes en modalidad semipresencial. La docencia a distancia me ha obligado a escribir gran parte de los comentarios que en su día recogí para los alumnos en varios archivos (noviembre y diciembre del 2007) y que ahora incluyo aquí, corregidos y mejorados, por si pudieran ser de utilidad para alguien.

Por su génesis, estos comentarios **no constituyen un material completo** de Estadística descriptiva. Pueden ser, si acaso, **un complemento** a material que sí abarque completamente los puntos que se quieran trabajar. Aquí se recogen sólo algunas –de las muchas posibles– dudas teóricas y prácticas que con más frecuencia tienen los alumnos a la hora de hacer este tipo de estadística. En general, las dudas que los alumnos puedan tener dependen tanto de las características de los materiales de la asignatura como de los alumnos (carácter, intereses, bagaje en ciencias). Las hay triviales y no tan triviales, pero ninguna «tonta».

Como indica el título, la estadística de estas hojas se limita al **análisis de una o dos variables**, que además han sido **medidas transversalmente**, en un instante de tiempo concreto. Ya «con los datos delante», una vez que se tienen claras las características de los datos y variables, la etapa del análisis descriptivo suele consistir en una recogida de la información en tablas y su expresión mediante gráficos y medidas numéricas, de modo que se aprecien mejor características que no es posible apreciar a simple vista de la muestra. Por este motivo se han dividido estos comentarios en las siguientes secciones: *Características de los datos*, *Construcción de tablas*, *Representación en gráficos* y *Cálculo de medidas numéricas*.

Índice

Una variable

Características de los datos

- Variables medidas transversal o longitudinalmente
- Variables estadísticas discretas y continuas
- Variables cualitativas
- Un poco más...
  - Variables estadísticas y variables aleatorias
  - Digresión sobre aparatos de medida y continuidad de las variables

Construcción de tablas

- Agrupación de los datos en clases
- Frecuencias absolutas y relativas
- Frecuencias relativas, en tanto por uno
- Frecuencias acumuladas

Representación en gráficos

- Histograma
- Diagrama de barras e histograma
- Un poco más...
  - Función de probabilidad e histogramas poblacional y muestral

Cálculo de medidas numéricas

- Marca de clase
- Cálculo de la varianza y la desviación típica
- Cuasivarianza (muestral)
- Tipificación de una variable
- Cálculo de las medidas de localización (incluida la mediana)
- Datos atípicos
- Distribuciones bimodales
- Cálculo de la moda
- Un poco más...
  - Interpretación de las fórmulas: asimetría, moda
  - MEDA
  - Coeficiente de apertura

Algunos ejercicios resueltos

- Ejercicio 1
- Ejercicio 2

Dos variables

Características de los datos

- Dos muestras univariantes y una muestra bivariante
- Vuelta a la estadística de una variable
  - Un poco más...
  - Datos multivariantes y datos multidimensionales
  - Vectores aleatorios

Construcción de tablas

- Distribuciones conjunta, marginales y condicionales

Representación en gráficos

- Diagrama de dispersión
- Recta de regresión
- Un poco más...
  - Ecuaciones de la recta
  - Representación a mano de una recta
  - Plantear una ecuación del enunciado
  - Resolución gráfica de algunos problemas
  - Otros tipos de regresión

Cálculo de medidas numéricas

- Correlación y relación no lineal entre variables
- Relación entre la correlación lineal y la recta de regresión
- Un poco más...
  - Inestabilidad numérica
  - Simulaciones
  - Apuntes sobre dependencia y correlación lineales

Algunos ejercicios resueltos

- Ejercicio 3
- Ejercicio 4

Una variable

Características de los datos

Variables medidas transversal o longitudinalmente

Hay una diferencia importante entre *datos transversales*, que recogen la información en un instante único y concreto de tiempo, y *datos longitudinales*, que lo hacen a lo largo del tiempo. (Aquí se entiende el término «longitudinal» en este sentido de medición a lo largo del tiempo, aunque por *datos longitudinales* se suele entender más generalmente los que se componen de funciones de variable posiblemente distinta al tiempo, que de hecho pueden estar medidas transversalmente en un instante de tiempo concreto.) En el primer caso se elige una muestra y se observan variables (una o más) en un instante de tiempo dado; en el segundo se elige una muestra y para cada elemento se observan variables (una o más) en varios momentos temporales distintos. Pero la distinción entre ambos tipos de datos no siempre está tan clara, ya que no tiene que ver únicamente con el momento en que se observan las variables sino con lo que las variables mismas significan. Por ejemplo, si la variable es «número de llamadas recibidas en un intervalo de tiempo», parece que la variable tiene cierto carácter longitudinal inherente; para complicar las cosas, esta variable puede a su vez medirse en una muestra en una sola vez, o puede hacerse un seguimiento de la variable en la muestra a lo largo del tiempo.

Es importante tener en cuenta de qué tipo son los datos a la hora de analizarlos: tanto a la hora de hacer el tratamiento, los gráficos y los cálculos, como de interpretarlos. Podríamos hacer la metáfora de compararlos, respectivamente, con una fotografía y un vídeo: si se trata de estudiar el efecto del paso del tiempo (movimiento), mientras la primera no informa de posibles movimientos a lo largo del tiempo, aunque en algunos casos pueda indicarlo, el vídeo sí muestra claramente si hay movimiento o no.

Dos formas de estudiar datos longitudinales que se suelen trabajar en cursos básicos de Estadística se basan, respectivamente, en las *series temporales* y en los *números índice*. Aunque los datos longitudinales tienen conceptos y métodos propios de análisis, esto no impide que algunos de los gráficos o medidas de los que se habla aquí tengan su utilidad con datos longitudinales. Por ejemplo, podemos representar un histograma de los errores después de ajustar un modelo a una serie temporal. Si una variable, aunque medida a lo largo del tiempo, tiende a tomar valores siempre en un conjunto acotado, también se puede representar el histograma de esos datos. A las series temporales, por ejemplo, cuando se les exige que sean *estacionarias*, lo que se garantiza es que, aunque los datos sean longitudinales, se les pueden aplicar con sentido ciertas operaciones (por ejemplo, las fórmulas de la media y de la varianza), porque se está uniendo información homogénea; sin la condición de estacionariedad las fórmulas se pueden aplicar numéricamente, pero dan información poco fiable.

En cualquier caso, la Estadística descriptiva de la que trata este documento es la correspondiente a una única variable medida transversalmente en una muestra de la población.

Variables estadísticas discretas y continuas

Las variables cuantitativas pueden ser discretas o continuas, dependiendo de si el conjunto de valores que pueden tomar es discreto o continuo.

Dentro de los *números reales* (que son los números con los que habitualmente trabajamos: 0, -3, 0,7, 2/9, $\sqrt{2}$, ...):

- Un subconjunto es discreto cuando sus valores se representan como puntos aislados en la recta de los reales y es continuo cuando al representarlos en la recta ocupan una franja continua (parte o toda la recta real), «sin huecos».
- Otro criterio equivalente es que un conjunto es discreto si se puede construir una sucesión con ellos, de forma que no se quede ninguno fuera (es un *conjunto numerable*), y es continuo si no se pueden ordenar en sucesión de manera que estén todos en la sucesión (es *no numerable*)

Lo anterior no tiene que ver, en principio, con que los valores de las variables tengan un número finito o infinito de decimales. La relación está en que en cuanto el conjunto de posibles valores incluye algún número irracional, hacen falta infinitos decimales –distintos y no periódicos– para poder representarlo. Los números irracionales, que están dentro de los reales, son los números que tienen infinitos decimales y los que hacen que los reales deje de ser un conjunto numerable. Los números que tienen un número finito o periódico de decimales son los números racionales. Por tanto, que una variable sea discreta o continua no equivale a que el número de decimales sea finito o no, respectivamente; la relación que puede y suele haber es que cuando puede tomar algún valor irracional son necesarios infinitos decimales. Pero es fácil definir una variable que sea discreta y necesite infinitos decimales: basta con que tome sus valores en el conjunto $\{0, \pi\}$.

Sin embargo, es verdad que frecuentemente sucederá que las variables discretas tomarán los valores en:

$\mathbb{N}$ = {0, 1, 2, 3...} (números naturales)

$\mathbb{Z}$ = {0, 1, -1, 2, -2...} (números enteros)

mientras que para las continuas consideraremos (al truncar los números) un número finito de decimales (porque en la práctica no podemos nunca considerar los infinitos, de esto se habla «filosóficamente» en el apartado Digresión sobre aparatos de medida y continuidad de las variables).

Variables cualitativas

Para este tipo de variables merece la pena hacer algunos comentarios que las prisas no suelen dejarnos hacer. Vamos a pensar conceptualmente en ideas sencillas que, sin embargo, son un buen fundamento para otras ideas más complicadas.

No se puede asignar de forma natural un valor numérico a cada valor cualitativo: esto implica que no se pueden ordenar de forma natural ni los valores de la muestra ni las clases en que se podrían agrupar estos valores (ordenamientos que sí induciría la asociación con números). Esto tiene como consecuencias:

- No tiene mucho sentido hablar de frecuencias acumuladas, aunque sí de frecuencias absolutas y relativas.
- No tiene sentido hablar de ciertos gráficos, como el histograma; aunque se utilizan otros como el diagrama de barras, el de sectores, pictogramas, etcétera.
- No tiene sentido hablar de ciertas medidas numéricas que dependen o de los valores o de un su ordenamiento, por ejemplo, la media, la varianza, las medidas de localización, etcétera; aunque sí se puede hablar de moda, porque se calcula a partir de las simples frecuencias.

Un poco más...

En esta sección se incluyen algunos comentarios de un nivel más alto, que pretenden por un lado facilitar la comprensión de lo visto y de los temas siguientes, y por otro satisfacer la curiosidad de quien quiera ir un poco más allá de lo estrictamente necesario según el temario.

Variables estadísticas y variables aleatorias

La Estadística trabaja –entre otros, pero principalmente– a dos niveles: el de los datos reales y el de la teoría (básicamente la Teoría de Probabilidades). En el primer nivel se estudian las *variables estadísticas*; en el segundo las *variables aleatorias*. Estas últimas, como su nombre indica, son variables que no toman valores de forma previsible, sino que cuando más se pueden determinar diciendo con qué probabilidad puede tomar cada uno de sus valores. La Estadística, como modo de trabajar, intenta asociar variables aleatorias a variables estadísticas; y viceversa. Se produce una equivalencia de conceptos entre los dos niveles: la frecuencia relativa de un nivel corresponde a la probabilidad en el otro, las medidas (media, varianza, asimetría, curtosis) de las variables estadísticas tienen sus análogas definidas para variables aleatorias, etcétera. He hecho el dibujo de abajo para representar esta dualidad.

Por ejemplo, si se están estudiando unos datos reales, se intenta ajustar un modelo teórico o variable aleatoria que explique/comparta características con la muestra, para así intentar comprender mejor el proceso que los ha generado. Por el contrario, también se puede recorrer el camino inverso, es decir, generar una muestra a partir de una variable aleatoria, lo que producirá datos que reflejan características del modelo: de una variable aleatoria simétrica (con la definición de *simetría* de las distribuciones de probabilidad), es de esperar que salga una muestra que tienda a ser simétrica (con la definición de *simetría* de las variables estadísticas), aunque no tiene por qué serlo exactamente, de hecho es difícil que lo sea. A medida que creciese el tamaño de la muestra, se haría más patente –por ejemplo, en el histograma– que la muestra comparte esa característica con el modelo del que procede.

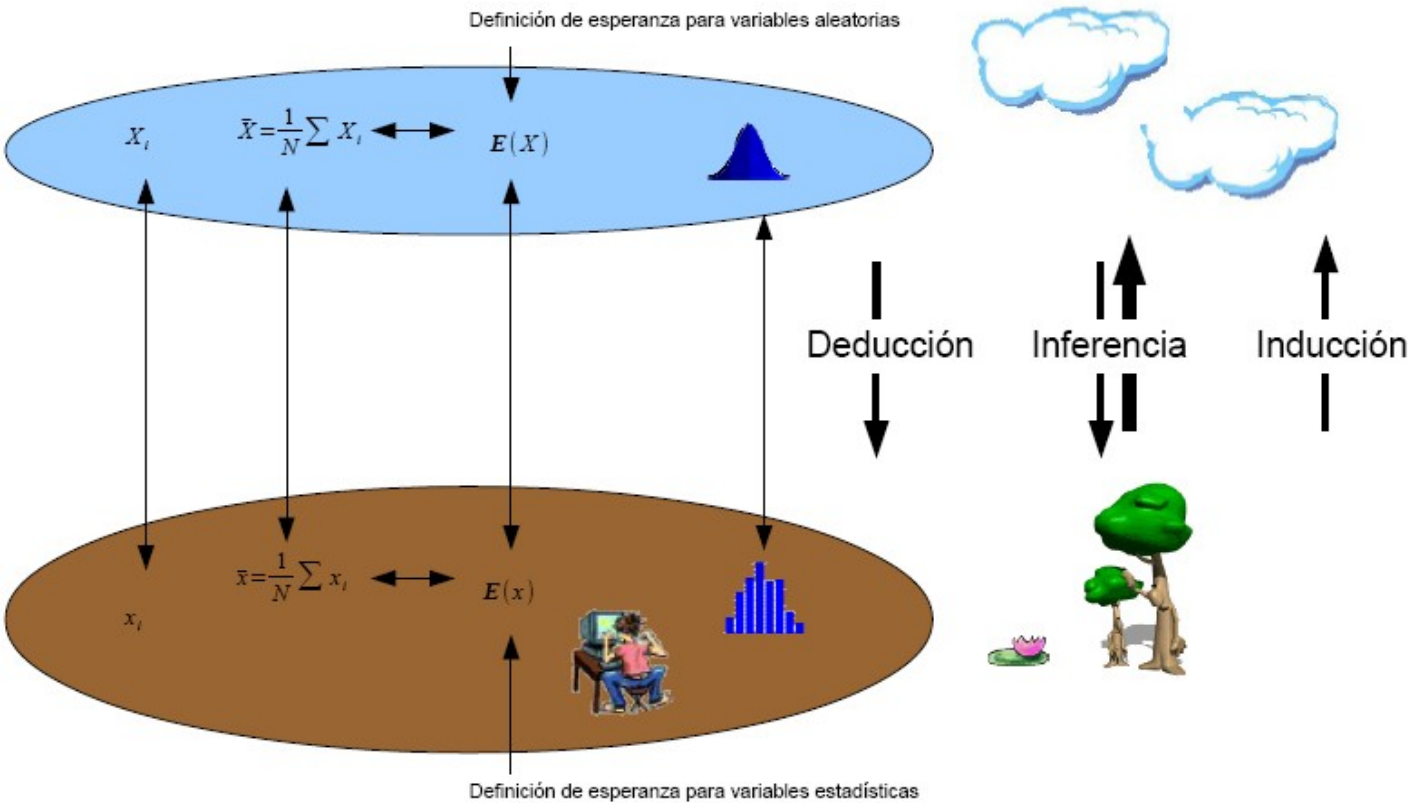


Figura: Representación de la dualidad datos-modelos

Digresión sobre aparatos de medida y continuidad de las variables

Quizá alguna persona dada a la filosofía se haya preguntado por qué se considera que la variable «peso de una persona» es cuantitativa continua, si no podemos dar el peso más que con un número finito de decimales, lo que implica un conjunto discreto de posibles valores de la variable. Este «filósofo de la ciencia» diría que esto se debe a que ningún aparato de medida que construyamos puede medir con infinitos decimales el peso de una persona y que, si lo hiciese, necesitaríamos un tiempo infinito para leer todas esas cifras... Esto es verdad (qué bonita es la Filosofía también). Sin embargo, lo explicado en el apartado Variables estadísticas y variables aleatorias puede ayudarnos a aclarar esta duda: aunque la variable estadística «peso» sea en último término siempre discreta (con más o menos posibles valores, dependiendo del aparato de medida y del tiempo que dediquemos a leer sus cifras), la variable aleatoria ideal «peso» se puede considerar continua, puesto que podemos imaginar personas de cualquier peso.

Ahora, el filósofo de la ciencia podría decirnos que la materia se compone en último término de partículas indivisibles con un peso mínimo determinado, y que por tanto hay un conjunto de valores discretos —y ningún otro intermedio— que debería poder tomar la variable «peso» teórica. Los físicos cuánticos le darían la razón, porque parece que la tiene.

Estamos hablando aquí de una variable estadística discreta y de dos posibles variables aleatorias con que modelizarla, una discreta y otra continua. El filósofo estaría defendiendo un modelo matemático discreto como más fiel a la realidad. Y lo es. Sin embargo, el modelo continuo queda legitimado por el hecho de que la escala de valores a la que trabajamos los aparatos de medida y nosotros está tan alejada de ese peso mínimo de las partículas elementales, que el error que se

comete por considerar el modelo continuo en vez de el discreto es despreciable frente al modelo matemático discreto. Por otro lado, no sólo el modelo continuo es válido sino que es el único camino posible (manejable) de manejar la información para levantarnos del sillón de filosofar y hacer Estadística...

Construcción de tablas

Agrupación de los datos en clases

Dada una muestra, hay veces en que no estamos interesados tanto en la información que aportan los valores concretos que ha tomado la variable estadística, sino que preferimos, necesitamos o nos conviene agrupar la información para hacerla más manejable y poder sacar información de ella grosso modo, sin que «los árboles nos impidan ver el bosque». De hecho, aunque son conceptos distintos que una variable sea discreta o continua y que esté agrupada o no, suele haber cierta confusión porque casi siempre se tiene que las variables discretas se dan sin agrupar y las continuas, como cada dato será distinto, agrupadas (para que las frecuencias absolutas no sean todas uno).

Clases

La forma de agrupar la información es considerar *clases*, que son subconjuntos contenidos dentro del conjunto total de posibles valores que puede tomar la variable estadística. Es importante que las clases sean disjuntas y que su unión sea el conjunto total, puesto que en otro caso habría valores que pertenecerían a varias clases o a ninguna, por lo que sería tenido en cuenta dos veces o se quedarían sin ser representado. Se suele decir en este caso que el conjunto de clases es una *partición* del conjunto de valores, puesto que lo dividimos en subconjuntos o clases. Veamos algunos ejemplos de clases:

1. Para una <u>variable cualitativa</u> se suele considerar que cada valor es una clase, pero quizá en algún caso puede ser interesante unir varios valores en una misma clase, por ejemplo: *iberia = {España, Portugal}*. Y hemos unido la información de dos valores de la variable «país» para formar la clase «iberia».

2. Para <u>variables cuantitativas discretas</u>, cada valor suele definir también una clase. No obstante, cuando los valores están muy dispersos y son poco frecuentes, conviene definir clases que incluyan a más de un valor. Un ejemplo sería: *clase1 = {0, 1, 2, 3}*, *clase2 = {4, 5, 6, 7}*... Es una forma de unir/resumir información, si no lo hacemos así tendríamos muchas clases y todas con frecuencias muy bajas.

3. Por último, para <u>variables cuantitativas continuas</u> las clases se definen mediante <u>intervalos</u> (por eso al hablar de histogramas utilizamos indistintamente estas dos palabras). En variables continuas es muy difícil («milagroso») que dos valores de la muestra sean iguales, así que aquí sí es una necesidad imperiosa lo de agrupar los datos, no como en los dos casos anteriores. Ejemplo: *clase1 = [0,5)*, *clase2 = [5,10)*, *clase3 = [10,15]*.

Agrupar los valores de la variable en clases es útil tanto para presentar los datos en tablas

como para representarlos gráficamente. Mientras que para las variables cualitativas el orden de las clases no tiene importancia, para las cuantitativas hay un orden natural (de menor a mayor) para ellas. Colocarlas en este orden es necesario tanto para facilitar la interpretación como para hacer algunos cálculos (las frecuencias acumuladas o las medidas basadas en las posiciones).

¿Cuántas clases considerar?

El problema de elegir el número de clases no tiene una solución única. Al agrupar los datos se pasa de saber los valores exactos a tener el rango de valores en que se mueve cada uno. Por este motivo tomar pocas clases hace que se pierda demasiada información, mientras que tomar demasiadas no ofrece ventajas con respecto a la presentación original de la muestra. Frecuentemente quien analiza los datos tiene que hacer varios dibujos con distinto número de clases, porque clases con demasiados valores no permiten ver los detalles y clases con pocos valores muestran tanto detalle que es como ver la muestra original sin agrupar.

En el caso de los intervalos, que es el que más nos interesa, la regla que se suele aplicar es considerar un número de ellos igual a $\sqrt{N}$, donde N es el tamaño de la muestra. Parece lógico que el número de clases dependa del tamaño de la muestra: a más datos, más clases.

Efecto de la partición

Dado que los intervalos (y las clases, en general) deben ser una partición completa de los posibles valores que puede tomar la variable (los haya tomado o no), es necesario incluir también los extremos. De hecho, si hubiese algún intervalo con frecuencia cero, lo incluiríamos tanto en la tabla como en los gráficos. Que haya un valor más en un intervalo que en los demás no es grave en estos casos, porque pocos elementos de la muestra tomarán ese valor.

Si tenemos dos particiones (un conjunto de clases por un lado y otro conjunto por otro), los resultados en ambos casos van a diferir poco si estas particiones son parecidas, pero serán muy distintos si las particiones lo son. Esto es así para todo lo que dependa de las clases, en concreto para tablas e histogramas. Es decir, si para cada intervalo de una partición tenemos que en la otra partición hay un intervalo que difiere muy poco de él, como los valores de la variable estadística que pueden haber caído en esa diferencia son pocos, los resultados serán muy parecidos. Por supuesto puede haber casos concretos en los que, si la zona en que difieren las particiones es de alta frecuencia, los resultados de ambas particiones diferirán mucho. No obstante, éstos son casos excepcionales.

Por tanto, cuando partimos el recorrido de posibles valores en intervalos, que el último intervalo sea cerrado por la derecha —en vez de abierto como los demás— no tiene ninguna consecuencia importante. La tendría si ese valor concreto tuviese una frecuencia alta, puesto que en ese caso esa clase de la partición sería distinta por construcción y habría que solucionarlo o tenerlo en cuenta a la hora de la interpretación.

Por ejemplo, en un ejercicio se pedía tomar como partición seis intervalos de igual longitud. Si se obliga a que los intervalos cubran la zona [3000, 26000], dado que 26000-3000 = 23000 no es múltiplo de 6, tenemos que los extremos de los seis intervalos tendrían que ser números con decimales; tampoco pasaría nada, se puede resolver así el ejercicio igual. Ahora bien, como según lo dicho antes los resultados no serían muy distintos por variar ligeramente la partición, podríamos dejar de cumplir el que la longitud tenga que ser la misma en todos los intervalos o el que se cubra la zona [3000, 26000]. En el primer caso podemos redondear los extremos de los intervalos a números cercanos (aunque a la hora de agrupar no aporta mucha ventaja), con lo que algún intervalo tendría que tener un poco más de longitud. En el segundo caso, se podría ampliar un poco la zona que

queremos cubrir, de manera que pudiésemos encontrar un partición de seis intervalos de igual longitud y cuyos extremos no sean números decimales que nos disgusten.

Información parcial

Cuando agrupamos la información de una muestra en clases, no debería ser necesario mencionar que no debemos nunca destruir los datos originales, porque puede ser necesario volver a ellos más adelante. Pero si los hemos destruido, o ya nos daban los datos agrupados, y los necesitamos, tendríamos que hacer suposiciones. Veamos algunos ejemplos:

- En el caso en que queramos información sobre algún intervalo que no coincide con ninguno de los utilizados para agrupar los datos, y no tengamos acceso a la muestra original, sino sólo a los datos agrupados, ¿qué podemos hacer? La solución que se suele adoptar es suponer que dentro de cada intervalo los datos estaban equidistribuidos antes de juntarlos y, en consecuencia, calcular la parte proporcional de los que estarían en la parte del intervalo que nos interesa. Esta idea es sencilla y no estará en general lejos de la verdad (sólo la podríamos conocer si conociésemos la muestra completa). Esto mismo es lo que se hace para calcular, en la misma situación, las medidas de localización.

- El concepto de *marca de clase*, que es el punto central o valor medio de los posibles valores del intervalo, hará de representante de todos los valores del intervalo, de modo que será el valor que se utilizará al hacer los cálculos. Obsérvese que con esto se está haciendo implícitamente la suposición de que todos los valores del intervalo habían tomado ese valor central en la muestra, antes de agruparlos; o, parecido, que se distribuían alrededor de él de modo que el valor central se puede considerar como su media. Obviamente, esto no es estrictamente verdad, pero el error en media no será grande, porque los verdaderos valores estarían muy probablemente por igual a uno y otro lado. En algunos cálculos de medidas, por ejemplo, se considera la marca de clase multiplicada por la frecuencia.

Intervalos en Matemáticas

Los corchetes indican que los valores se incluyen, mientras que los paréntesis indican que no. Es decir, si estamos hablando de una variable continua:

$[a, b]$ ≡ Todos los valores entre a y b, ambos incluidos

$[a, b)$ ≡ Todos los valores entre a y b, incluido el a pero no el b

$(a, b]$ ≡ Todos los valores entre a y b, incluido el b pero no el a

(a, b) ≡ Todos los valores entre a y b, sin incluir ni el a ni el b

Al agrupar datos suelen considerarse intervalos cerrados por la izquierda y abiertos por la derecha, salvo el último, que se cierra por la derecha para no dejar ese valor fuera de la partición. Otras veces se toma $(-\infty, b]$ como intervalo inferior o $[a, +\infty)$ como intervalo superior, para que recojan todos los valores extremos.

Frecuencias absolutas y relativas

Las frecuencias absolutas no aportan una información completa si no se conoce el tamaño de la muestra, porque no es lo mismo que algo suceda 3 veces de 10 que de 500. Por otro lado, conocer las frecuencias relativas, aunque facilita el pensar, por estar en tanto por uno, tampoco es suficiente: una frecuencia relativa de 0,5 puede provenir de 1 vez de 2 o de 15 de 30; y la interpretación obviamente no debe ser la misma. Es decir, conocer el tamaño de la muestra es imprescindible en todo estudio estadístico serio. Si N es el tamaño de la muestra de nuestra variable estadística, este parámetro nos permite tanto ir de las frecuencias absolutas a las relativas como recorrer el camino inverso.

Frecuencias relativas, en tanto por uno

Hay veces en que se utiliza el tanto por ciento, en vez del tanto por uno, para la frecuencia relativa. No hace falta decir, por obvio, que la interpretación es la misma y que se pueden convertir fácilmente los datos en tanto por uno dividiéndolos por cien. La ventaja de esta unidad es que en nuestra vida diaria estamos más acostumbrados a utilizar el tanto por ciento que el tanto por uno. La ventaja del tanto por uno es que está más cerca del concepto de probabilidad que se verá más adelante y que se menciona en el apartado Variables estadísticas y variables aleatorias. Por motivos históricos la escala que se utiliza para la probabilidad es el tanto por uno (podría ser otra). La diferencia entre utilizar unas unidades u otras es sólo una diferencia de proporción (entre dibujos, longitudes, etcétera), no de forma. No obstante, la definición propia de frecuencia relativa implica utilizar el tanto por uno.

Frecuencias acumuladas

Como el nombre indica, la frecuencia acumulada para cada clase es el número que resulta de sumar la frecuencia de su clase y de las anteriores. Es, por tanto, no decreciente: crece o, si la frecuencia de la clase anterior es cero, permanece igual.

De esta definición se deduce que para variables cualitativas, dado que no hay un orden natural de las clases, no es un concepto que tenga valor. Estas frecuencias tienen sentido completo sólo para variables cuantitativas.

Hay una *frecuencia absoluta acumulada* y una *frecuencia relativa acumulada*, no hay que confundirlas. Es necesario que aprendamos cómo relacionar la frecuencia acumulada con la absoluta y la relativa; esto es sencillo a partir de sus definiciones:

- Por un lado, la relación que hay entre las frecuencias absoluta y relativa será la misma que entre las frecuencias absoluta acumulada y relativa acumulada: se pasa de la primera a la segunda dividiendo por el tamaño de la muestra, y de la segunda a la primera multiplicando por ese tamaño.
- Por otro lado, para obtener las frecuencias absolutas, n_i, a partir de sus acumuladas, N_i, observemos primero cómo se han construido las acumuladas:

$$\left.\begin{array}{c} N_1=n_1 \\ N_2=n_1+n_2 \\ N_3=n_1+n_2+n_3 \\ \vdots \end{array}\right\} \rightarrow \left\{\begin{array}{c} n_1=N_1 \\ n_2=N_2-N_1 \\ n_3=N_3-N_2 \\ \vdots \end{array}\right.$$

La importancia de estas frecuencias es que son una herramienta clave para calcular algunas medidas de las variables estadísticas, en concreto las medidas de localización y la mediana (que es a la vez medida de centralización y de localización). Como con las otras frecuencias, podemos expresarlas en forma de tabla o gráficamente. En este último caso esto nos llevaría a un histograma con sus barras en orden no decreciente si se observan de izquierda a derecha. La interpretación de cada barra es que representa la frecuencia de que suceda que la variable estadística tome un valor en ese intervalo o en los anteriores. Nótese que dado que las medidas de localización utilizan el orden de los datos en la muestra (cuando la ordenamos de menor a mayor), las frecuencias acumuladas sirven para ver rápidamente en qué intervalo está la observación de determinada posición, que es lo que necesitamos en primer lugar para calcular estas medidas. En el apartado Cálculo de las medidas de localización (y la mediana) se ve qué se hace una vez que se identifica el intervalo en el que está la medida que nos interesa.

Representación en gráficos

Histograma

Es una de las representaciones gráficas más utilizada en la Estadística.

Área

Cada clase (por tanto se utiliza para datos agrupados) se representa por un rectángulo cuya área es igual a la frecuencia relativa de la clase.

Altura

De lo anterior se deduce que la altura de cada rectángulo debe ser igual a la frecuencia relativa dividida por la amplitud o longitud del intervalo. También se deduce que el área total encerrada por todas las barras es igual a uno (el máximo de la escala en que se mide la probabilidad).

Longitud

Puede suceder que la longitud de todos los intervalos no sea la misma, como sugeriría, por ejemplo, alguna transformación que se ha aplicado a los datos o el hecho de que los valores que toma la variable están más separados cuanto mayores son. Pero en general se suelen considerar todos los intervalos de la misma longitud, por lo que en este caso las alturas de las barras son proporcionales a las frecuencias absolutas (y, por tanto, también a las relativas).

Respecto a la longitud numérica concreta de los intervalos, lo que interesa es que todos los

datos de la muestra aparezcan en el histograma, es decir, que todas las clases cubran el rango de valores de la muestra (del dato menor al mayor). Por esto, cuando la longitud se toma igual para todos los intervalos se calcula dividiendo este rango entre el número de clases, $\sqrt{N}$, que se ha sugerido en el apartado Agrupación de los datos en clases. Para ver el efecto de considerar distintos números de clases se presenta un gráfico del libro *Introducción a la Estadística para las Ciencias Sociales*, de Daniel Peña y Juan Romo.

No obstante, no todos los programas de ordenador trabajan igual ni siguen las mismas reglas, así que el histograma dependerá del programa que lo haya dibujado. La idea importante es que aunque la longitud de los bloques (o su número) influye relativamente poco para pequeños cambios, sí lo hace para los grandes. Así, dos personas pueden hacer histogramas con intervalos ligeramente distintos y ambos estarán bien, porque lo importante no es tanto el dibujo sino que de él se pueda sacar información válida. De esto también se ha hablado en el apartado Agrupación de los datos en clases. En el Ejercicio 2 se pueden ver algunos ejemplos más de histograma, correspondientes a una distribución simétrica y dos asimétricas.

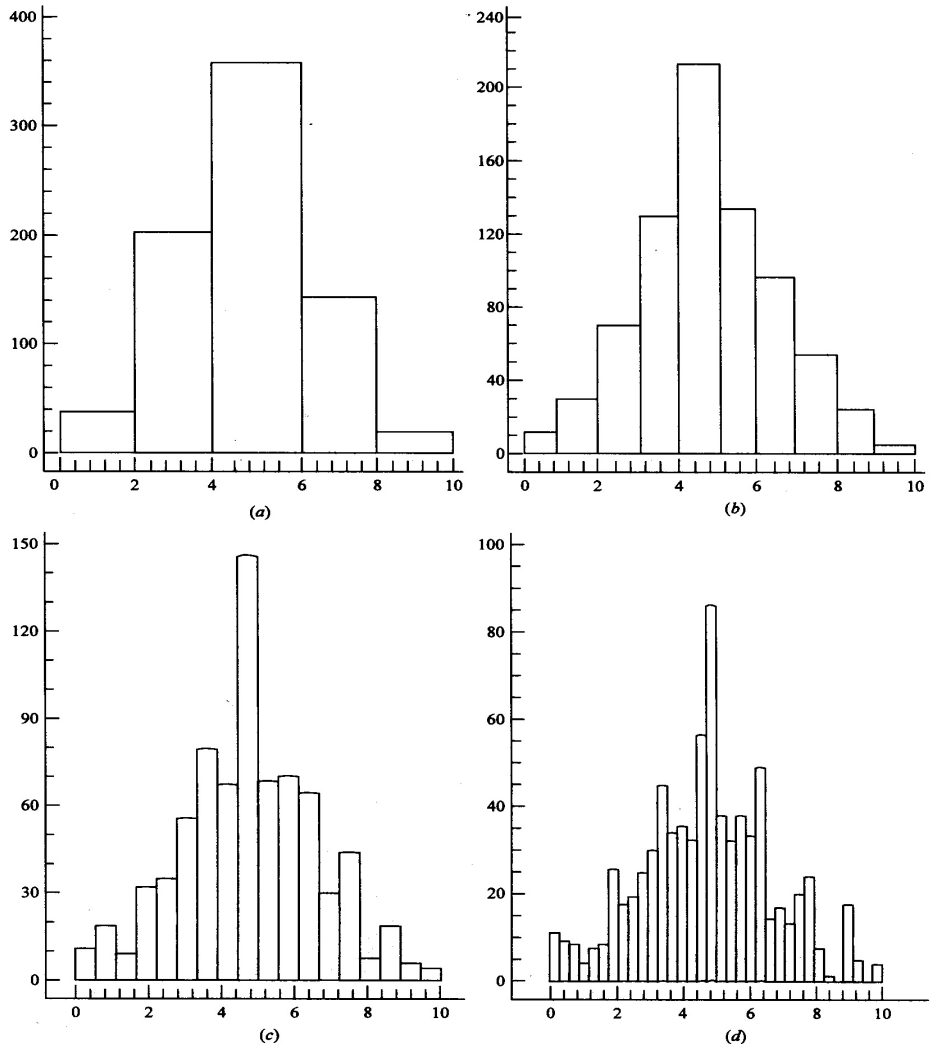


Figura 3.6. *Histograma para la variable NOTAS con diferente número de clases: (a) 5; (b) 10; (c) 18; (d) 34.*

Densidad

La información de la frecuencia de un intervalo no se puede interpretar bien sin tener en cuenta la amplitud o longitud del intervalo; de ahí que se defina el concepto de *densidad*, resultado de dividir la frecuencia por la amplitud. La densidad coincide con la altura. Cuando esta amplitud de todos los intervalos es la misma, como suele ser, las densidades son proporcionales a las frecuencias.

Otras versiones del histograma

Existen algunas otras versiones, ligeramente distintas, del histograma. Todas cumplen con el objetivo de mostrar la distribución de los valores de una muestra, pero es importante a la hora de la interpretación tener claro qué versión se tiene delante.

Hay versiones en las que la amplitud de los intervalos es siempre la misma, lo que hará que las alturas y las densidades sean proporcionales a las frecuencias.

También se puede representar el histograma de manera que el área que sumen todas las barras sea N (tamaño muestral). Para ello, se hace que cada barra tenga un área igual a su frecuencia absoluta. A su vez, como el área de un rectángulo es base por altura, la altura se toma como la frecuencia absoluta dividida por la longitud del intervalo; ésta es la fórmula de la densidad: n_i/l_i. Cuando se representaba el histograma de manera que el área que sumaban las barras era uno, cada barra tendrá un área igual a la frecuencia relativa de la clase, y la densidad se calculará dividiendo esta frecuencias por la longitud del intervalo: f_i/l_i.

Estos dos tipos de histograma son proporcionales entre sí, porque $f_i=n_i/N$, y ambos son válidos para sacar información cualitativa.

Diagrama de barras e histograma

Una diferencia entre el histograma y el diagrama de barras es que en el primero las barras se dibujan juntas, para manifestar la continuidad de los intervalos. Otra diferencia es que en un diagrama de barras, que suele utilizarse para variables cualitativas, en realidad el orden de las barras suele dar igual, mientras que en el histograma viene impuesto por el orden de los intervalos.

En general se puede decir en Matemáticas que todo lo que sea verdad se puede hacer: no hay verdades prohibidas, con las definiciones y las reglas en la mano, se puede avanzar y si no se quebrantan las reglas lo que se obtiene será verdad. Dentro de lo que se puede hacer, hay cosas que tiene sentido hacer y cosas que no: una definición de algo que no existe en realidad, por ejemplo, se puede proponer pero no se debe; o una definición que sólo utiliza uno mismo, tampoco tiene mucho sentido. Hay ventajas en ponernos todos de acuerdo en ciertas cosas, aun habiendo libertad. Si el concepto es importante, debe quedar claramente definido. Si no lo es tanto, como todos estos gráficos sencillos, se deja cierta libertad, siempre que la información que se transmite se pueda interpretar por los demás sin mucho esfuerzo. Dibujar las barras juntas o separadas, por ejemplo, es una cuestión de estilo más que de Matemática. Veamos las ideas que hay detrás de cada uno de estos tipos de gráfico, para saber cuál utilizar en cada caso:

Diagrama de barras

1. Se utiliza principalmente para variables cualitativas (no hay muchos tipos de gráfico para estas variables). En este caso, el orden de las barras no importa para la interpretación.
2. Utiliza una barra para representar cada clase.
3. La altura de la barra debe ser la frecuencia (absoluta o relativa, sólo variará la proporción del dibujo, pero no su forma).
4. Las barras se representan con una ligera separación entre ellas, puesto que la idea es construir una barra encima de cada clase y juntarlas dificulta su visión.
5. La anchura de las barras no suele aportar información (no representa nada), no cambia de valor si la muestra tiene más datos.

Histograma

1. Cada rectángulo debe tener un área igual a la frecuencia de la clase (absoluta o relativa, sólo variará la proporción del dibujo, no su forma).
2. La altura debe ser igual al área dividida por la longitud del intervalo de la clase.
3. La longitud de los intervalos podría ser distinta en algunos casos, pero en la mayoría va a ser la misma para todas las clases. La longitud se elegirá dividiendo el rango de valores que queremos cubrir por el número de intervalos que queremos tomar (hay una regla sugerida para agrupar los datos, ver el apartado Agrupación de los datos en clases). Entonces, la longitud de los intervalos —y la anchura de las barras— disminuye a medida que aumenta el tamaño de la muestra.
4. Cuando la variable es continua, como los intervalos suelen estar seguidos, conviene representar las barras juntas, para no pensar que hay entre medias valores con frecuencias nulas.

Como se ve, en uno la anchura de las barras depende del tamaño de la muestra y en el otro no, el uno quiere representar frecuencias por la altura y el otro por el área, etcétera. Ahora bien, si nos empecinamos y en el diagrama de barras en vez de variables cualitativas representamos cuantitativas, obviamente habrá ahora un orden natural entre las clases y las barras, que no se podrán colocar de cualquier forma. Si las clases están separadas por definición (por ser variables discretas), entonces conviene dibujar las barras separadas. Si son variables continuas, los intervalos suelen estar seguidos, así que conviene representarlos seguidos. Si representamos la frecuencia en la altura y tomamos las barras de la misma anchura (como este tipo de gráficos suele hacer), entonces se queda sin representar la longitud de los intervalos, cosa mala... Parece que no aporta nada forzar las cosas. Cuando tenemos una variable cuantitativa discreta, a medida que aumentásemos el tamaño de la muestra sólo las clases que quedan encima de los valores enteros tendrían frecuencia no nula, mientras que las demás clases desaparecerían del dibujo, dando la sensación de que se trata de un diagrama de barras y de que las barras están separadas, cuando en realidad lo que ha sucedido es que hay barras que no se ven por tener altura cero.

En resumen, lo mejor es utilizar el diagrama de barras para variables cualitativas y el histograma para las cuantitativas, aunque cuando éstas son discretas y las longitudes de los intervalos menores que uno aparezca un histograma que se parece a un diagrama de barras.

En el apartado Histograma hay un ejemplo de este tipo de gráfico; el siguiente es un ejemplo de diagrama de barras.

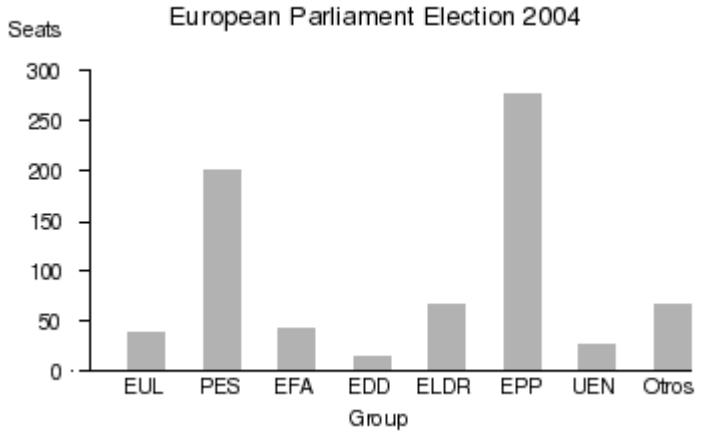


Figura: Ejemplo de diagrama de barras

Un poco más...

En esta sección se incluyen algunos comentarios de un nivel más alto, que pretenden por un lado facilitar la comprensión de lo visto y de los temas siguientes, y por otro satisfacer la curiosidad de quien quiera ir un poco más allá de lo estrictamente necesario según el temario.

Función de probabilidad e histogramas poblacional y muestral

Función de probabilidad: función de masa y función de densidad

A una variable aleatoria se le asocia una *función de probabilidad*, que dice con qué probabilidad la variable puede tomar cada valor. Si la variable aleatoria es discreta esta función de llama *función de masa* y consiste en asignar un número (probabilidad) a cada posible valor de la variable; si la variable es continua, se da la *función de densidad*, que es una función continua que asigna, a su modo, probabilidad a los infinitos posibles valores de la variable.

Histograma poblacional

Para un intervalo, *I*, de posibles valores de una variable aleatoria, *X*, utilizando la función de probabilidad se puede calcular la probabilidad siguiente $p := P(X \in I)$. (El concepto de *probabilidad* se explica en la Teoría de la Probabilidad, aunque todos tenemos una idea intuitiva de él.) Ahora, dado un tamaño muestral *N*, se puede estimar el número de veces de entre *N* que la variable aleatoria tomaría un valor del intervalo: para hacer este recuento se puede considerar la variable auxiliar $Y \sim B(N, p)$, que se va a «encargar de contar». La cantidad que se quiere hallar es la *frecuencia absoluta esperada* y se calcularía como $n_I = N \cdot p = N \cdot P(X \in I)$, de donde su *frecuencia relativa esperada* sería $f_I = \frac{1}{N} \cdot N \cdot p = p = P(X \in I)$. Nótese que la distribución de *Y* es siempre, por definición, binomial; no se ha supuesto ninguna distribución concreta para *X;* la distribución de *X* tiene influencia indirecta en la de *Y* sólo a través del parámetro *p*. En resumen:

- Variable original: X
- Intervalo: I

- Probabilidad del suceso $\{X \in I\}$: $p = P(X \in I)$
- Tamaño muestral: N
- Variable binomial auxiliar (para calcular cuántas veces se caería en *I*): $Y \sim B(N, p)$
- Esperanza o valor medio de *Y*: $N \cdot p = N \cdot P(X \in I)$
- Frecuencia absoluta esperada del intervalo *I*: $n_I = N \cdot p = N \cdot P(X \in I)$
- Frecuencia relativa esperada del intervalo *I*: $f_I = \frac{1}{N} \cdot n_I = \frac{1}{N} \cdot N \cdot p = p = P(X \in I)$

Para una partición completa y disjunta del conjunto de posibles valores de la variable (se ve en el apartado Agrupación de los datos en clases), es posible hacer esto para cada uno de los intervalos, de manera que si se dibuja un histograma con estas frecuencias esperadas relativas como alturas de las barras, tendremos lo que se podría llamar *histograma teórico, esperado* o *poblacional* (esta nominación es mía, no he visto este concepto no estos nombres en la literatura, lo que no quiera decir que no existan...). Nótese que las barras del histograma sumarían un área igual a uno, porque sumando para todos los intervalos la frecuencia relativa esperada: $\sum_i f_{I_i} = \sum_i P(X \in I_i) = 1$.

Otra forma de hacer los cálculos de todos los intervalos a la vez sería considerando una variable aleatoria *Y* con *distribución multinomial* (en vez de la binomial anterior) de parámetros $p_i := P(X \in I_i)$.

Histograma muestral

En el «nivel de los datos», a una variable estadística se le asocia el histograma calculado con las frecuencias relativas reales de la muestra, como se describe en el apartado Histograma. Este concepto se define para la variable estadística, mientras que los dos anteriores se definían para la variable aleatoria.

No se suele hablar de la distinción entre los dos tipos de histograma anteriores, lo que puede resultar confuso en ciertas ocasiones. Hay que distinguir estos tres conceptos, es decir, hay que tener claro que una cosa es el histograma de unos datos, otra es el histograma teórico y una tercera es la función de probabilidad teórica. Cuando la muestra crece, es decir, *N* crece, los dos tipos de histogramas anteriores se parecerán más y más a la función de probabilidad –se dice entonces que la *estiman*– a la vez que se parecerán también más entre sí.

Cálculo de medidas numéricas

Marca de clase

Cuando los datos están ya agrupados no se dispone de los valores exactos que toma la variable estadística. Sin embargo, hay ocasiones en que querríamos tener esos valores para utilizarlos en los cálculos. La *marca de clase*, que en el caso de los intervalos es su valor central, es como el «representante» de todos los valores del intervalo.

La idea se puede explicar pensando en reconstruir la muestra, o, mejor dicho, construir una

muestra ficticia tal que si se agrupa da lugar a los mismos datos agrupados de que disponemos. Para construir esta muestra, se supone que todos los datos de cada intervalo tomaban el valor de su marca de clase; el error que se comete haciendo esta suposición es, generalmente y en media, nulo. Por ejemplo: si en el intervalo [1,3) hay cinco datos (frecuencia absoluta), al reconstruir la muestra, cosa que no es necesario hacer explícitamente, se habla de ella aquí para comprender el concepto, pondríamos:

...,2, 2, 2, 2, 2,...

haciendo lo mismo con todas las demás frecuencias.

Una vez hecho lo anterior, podríamos aplicar la fórmula de datos sin agrupar a esta muestra ficticia, o, equivalentemente, aplicar las fórmulas que se dan para datos ya agrupados (que no son más que las mismas fórmulas pero expresadas ya en función de las marcas de clase y las frecuencias absolutas).

Cálculo de la varianza y la desviación típica

Para calcular la desviación típica es necesario calcular primero la varianza, porque se define en función de ella. Ahora, para calcular la varianza se pueden utilizar varias fórmulas: una es la fórmula de su definición y otras, a veces más prácticas, según estén los datos, se derivan de ella:

$$Var(X)=s_X^2=\frac{1}{m}\sum(x_i-\bar{x})^2=\cdots=\frac{1}{m}\sum x_i^2-\left(\frac{1}{m}\sum x_i\right)^2=\frac{1}{m}\sum x_i^2-(\bar{x})^2.$$

A la hora de hacer los cálculos, es importante ir preparando con orden en la tabla los datos que se van a necesitar después en las fórmulas: esto evita equivocarse.

Una observación a la hora de hacer estos cálculos es la siguiente. Supongamos que tenemos que calcular las cantidades $(x_i-\bar{x})^2$; aunque es cierto que la fórmula $(a-b)^2=a^2-2\text{ab}+b^2$ haría que pudiésemos hacer $(x_i-\bar{x})^2=x_i^2-2x_i\bar{x}+\bar{x}^2$ y calcular así cada valor; esto no es muy útil, porque es más cómodo incluir en la tabla las cantidades $(x_i-\bar{x})$ y después sus cuadrados.

Cuasivarianza (muestral)

La variabilidad es un concepto que cuantifica lo cerca que los datos de una muestra se sitúan de la medida de centralización que se esté considerando. La varianza, una de las formas de medir la variabilidad, calcula cuánto se ha separado cada dato de la media, una de las formas de medir la centralización; después quita el signo elevando al cuadrado cada una de estas cantidades y finalmente hace la media aritmética de todas estas últimas cantidades. En esta media aritmética se suman todas las cantidades y se divide entre el número de ellas:

$$s^2=\sum_{i=1}^{N}\frac{(x_i-\bar{x})^2}{N}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2$$

donde aquí cada x_i es el dato directamente tomado de la muestra, que tiene N datos. Esta fórmula puede escribirse de otra forma cuando los datos están agrupados por valores o en clases; hay que tener cuidado porque entonces quizá se utiliza x_i para representar la marca de clase. Una indicación práctica para hacer esos cálculos es que, tanto si los hacemos a mano como con el ordenador, conviene dividir por N al sumatorio final, no a cada término del sumatorio; es decir, conviene aplicar la última expresión, no la penúltima.

Sucede que la varianza no es todo lo buena que se desearía, no es *insesgada*, no se aproxima al valor esperado (la *varianza poblacional* que se define en la Teoría de la Probabilidad). Entonces se corrige la fórmula anterior y se define la *cuasivarianza*, cuya fórmula es:

$$\tilde{s}^2=\sum_{i=1}^{N}\frac{(x_i-\bar{x})^2}{N-1}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2.$$

La *cuasidesviación típica muestral* se define como $\sqrt{\tilde{s}^2}$. Se cumple la relación siguiente (se supone, obviamente, que $N > 1$):

$$\begin{cases} s^2 N=\sum_{i=1}^{N}(x_i-\bar{x})^2 \\ \tilde{s}^2(N-1)=\sum_{i=1}^{N}(x_i-\bar{x})^2 \end{cases} \rightarrow \quad s^2 N=\tilde{s}^2(N-1) \quad \rightarrow \quad s^2\frac{N}{N-1}=\tilde{s}^2.$$

Y como $\frac{N}{N-1}\approx 1$ para valores de N grandes, pero para valores pequeños de N no se cumple esta aproximación, la cuasivarianza es mejor para valores pequeños de N y es igual que la varianza para valores grandes de N. La cuasivarianza, para estos valores pequeños de N se aproxima más al valor real de la varianza, porque es un estimador *insesgado*. Por esto es la que se utiliza en la práctica, es la que utilizan la mayoría de los programas y lenguajes de programación.

Hay que tener en cuenta qué fórmula se está usando cuando se hacen los cálculos a mano o con el ordenador. Para complicar un poco más las cosas, los programas de ordenador suelen llamar «varianza» a la cuasivarianza, porque suponen que el usuario conoce la diferencia o, si no la conoce, le da igual; no obstante, con un ejemplo de pocos datos concretos que introduzcamos, o consultando la ayuda del programa, se puede descubrir qué fórmula aplica.

En estos conceptos anteriores a veces se añade el adjetivo «muestral» para indicar que son los conceptos que se definen en «el nivel de las variables estadísticas» del que se habla en el apartado Variables estadísticas y variables aleatorias; en «el nivel de las variables aleatorias» existen los conceptos poblacionales/teóricos equivalentes, que se llamarán *media poblacional*, *varianza poblacional* y *desviación típica poblacional*.

Tipificación de una variable

Frecuentemente surge la duda de cómo comparar dos datos cuando cada uno pertenece a una muestra distinta. No se pueden comparar sus valores sin más, antes hay que transformarlos.

Rápidamente viene a la cabeza la idea de que el coeficiente de variación es una medida de dispersión sin unidades de toda una muestra, lo que hace que sirva para comparar la dispersión de dos

muestras enteras distintas. Ésta es la idea, aunque algo más hay que hacer para comparar dos datos concretos, puesto que el coeficiente de variación no está dado para un dato concreto, sino para toda una muestra.

La idea es comparar cada dato con la media de su muestra, pero teniendo en cuenta a la vez la variabilidad de los datos de su muestra. Por ejemplo, para la variable «nota académica», no es lo mismo que un valor esté más o menos cerca de su media, como tampoco es lo mismo que todos los demás alumnos hayan sacado o no notas parecidas. Finalmente, para poder comparar datos de distintos grupos, queremos que los valores transformados no tengan unidades. Es decir, queremos:

- Dentro de cada grupo:
  - Tener en cuenta la distancia a la media.
  - Tener en cuenta la variabilidad del grupo.
- Entre grupos:
  - Que la medida transformada no tenga unidades

Los anteriores objetivos los cumple la transformación que se llama *tipificar*. Un dato de una muestra se tipifica restándole la media de su grupo y dividiendo el resultado por la desviación típica del grupo. Esto es, cada valor x_i de la muestra se tipifica haciendo:

$$z_i = \frac{x_i - \bar{x}}{\sigma}.$$

Esta fórmula es un cociente de dos cantidades. La de arriba es la distancia a la media, y la de abajo es la referencia que normaliza, es decir, ahora z_i nos está diciendo cuántas veces, medido «con una barra» de longitud igual a la desviación típica, el dato x_i se aleja de la media de su muestra. Podemos pensar que medimos la distancia del dato a la media «utilizando como patrón o unidad de medida la desviación típica». Con lo que se ha explicado de las unidades en el apartado Coeficiente de apertura se puede comprobar fácilmente que esta cantidad no tiene unidades, porque las tres cantidades que aparecen en la fórmula están en las mismas unidades y, por tanto, se cancelan entre sí.

Cálculo de las medidas de localización (incluida la mediana)

Las medidas de localización dan una idea de cómo se sitúan los datos de la muestra. Para ello dan valores que dividen la muestra en dos partes de proporciones dadas: por ejemplo, dan un valor que divide la muestra en una mitad con los datos de menor valor y otra mitad con los de mayor valor, o dan un valor que está entre el 10% de los datos de menor valor y el 90% restante. Esto es lo importante de estas medidas, la idea de lo que pretenden hacer, porque en la práctica las cosas son un poco más difíciles. No hay una definición única, por ejemplo, de los *cuantiles*/*percentiles* (se habla de ellos porque incluyen a los deciles, cuartiles y mediana). Una muestra de esta falta de unicidad es que en la ayuda de la función del lenguaje de programación *R* que calcula los cuantiles se informa de que se puede elegir para el cálculo uno de los ¡nueve! algoritmos/definiciones distintos: tres para variables discretas y seis para continuas. Dependiendo de la fuente que tomemos, podemos leer definiciones ligeramente diferentes, todas ellas válidas.

Una primera cosa que se debería aclarar en la definición de estas medidas es si pueden tomar sólo valores de los que aparecen en la muestra o también algún valor intermedio. Hay libros en que se dice que el percentil de orden *p* es «el menor valor superior al p% de los datos», pero no siempre

queda claro si se refiere a valores de los de la muestra o de los que podría haber tomado la variable estadística. En realidad esto no es muy importante, ya que la información se transmite igual. Si nos dicen que el primer cuartil es 15,5 y sabemos que eso significa que el 25% de los datos de la muestra están por debajo de esa cantidad. En general nosotros vamos a permitir que estas medidas pueden tomar valores distintos a los de la muestra. Antes de presentar qué procedimientos se siguen en los casos de datos sin agrupar y agrupados (los discretos suelen estar sin agrupar y los continuos, por necesidad, agrupados), algo que tiene que quedar claro es que una cosa es hablar de la posición de un valor de la variable en la muestra y otra es hablar del valor mismo de ese dato. Vamos a llamar N al tamaño de la muestra.

Datos sin agrupar

Para valores sin agrupar no hay mucha dificultad. El procedimiento es el siguiente:

1) Ordenamos la muestra de menor a mayor.

2) Queremos encontrar el valor que deja por debajo de él $\frac{k \cdot N}{100}$ observaciones (podemos pensar que este número que buscamos indica *número de observaciones* o *posición*, da igual). Puede suceder:

   - $\frac{k \cdot N}{100} \notin \mathbb{N}$, es decir, que esta cantidad no sea un número natural (no puede indicar una posición): Entonces redondeamos este valor hacia arriba, esto es, tomamos el menor número natural mayor que esa cantidad; esto es equivalente a considerar la parte entera de esa cantidad (quitarle los decimales) y sumarle uno. Entre redondear hacia abajo y obtener un valor que deja menos proporción de datos de la deseada, o redondear hacia arriba y pasarnos, preferimos esto segundo. Ya tenemos el número que nos indica la posición. Vamos a la muestra ordenada y buscamos el valor de la variable en esa posición. Tomamos ese valor como cuantil.

   - $\frac{k \cdot N}{100} \in \mathbb{N}$, es decir, da la casualidad de que esta cantidad es un número natural: Entonces no hace falta redondear y vamos a coger dos posiciones: la de ese número y la siguiente. Por convenio (uno de los posibles) es frecuente considerar como cuantil la semisuma de los valores que están en esas dos posiciones, es decir, la media de esos dos valores de la muestra ordenada.

Datos agrupados

Para valores agrupados el procedimiento es algo distinto.

1) Ordenamos las clases (intervalos, generalmente) de menor a mayor. Obtenemos las frecuencias absolutas acumuladas, para después localizar rápidamente el intervalo donde está la posición que nos interesa.

2) Aunque nuestro objetivo es buscar un valor que deje por debajo k% de los datos, como los datos están agrupados sólo podemos encontrar el intervalo al que pertenece el valor $\frac{k \cdot N}{100}$. Identificamos este intervalo porque la frecuencia absoluta acumulada del intervalo anterior es

mejor que esa cantidad y la del intervalo posterior es ya mayor. En este caso nos da igual que $\frac{k \cdot N}{100}$ sea un número natural o no (no necesitamos redondear), porque no es necesario que indique posición.

3) Una vez que está identificada la clase en la que está la medida de localización que buscamos, tenemos que decidir qué valor de todo ese intervalo coger para la medida. La idea que vamos a seguir aquí (un posible criterio, de entre varios posibles) es coger el valor que esté a una distancia, del extremo inferior del intervalo, proporcional a la distancia que hay entre $\frac{k \cdot N}{100}$ y la frecuencia absoluta del intervalo anterior al elegido.

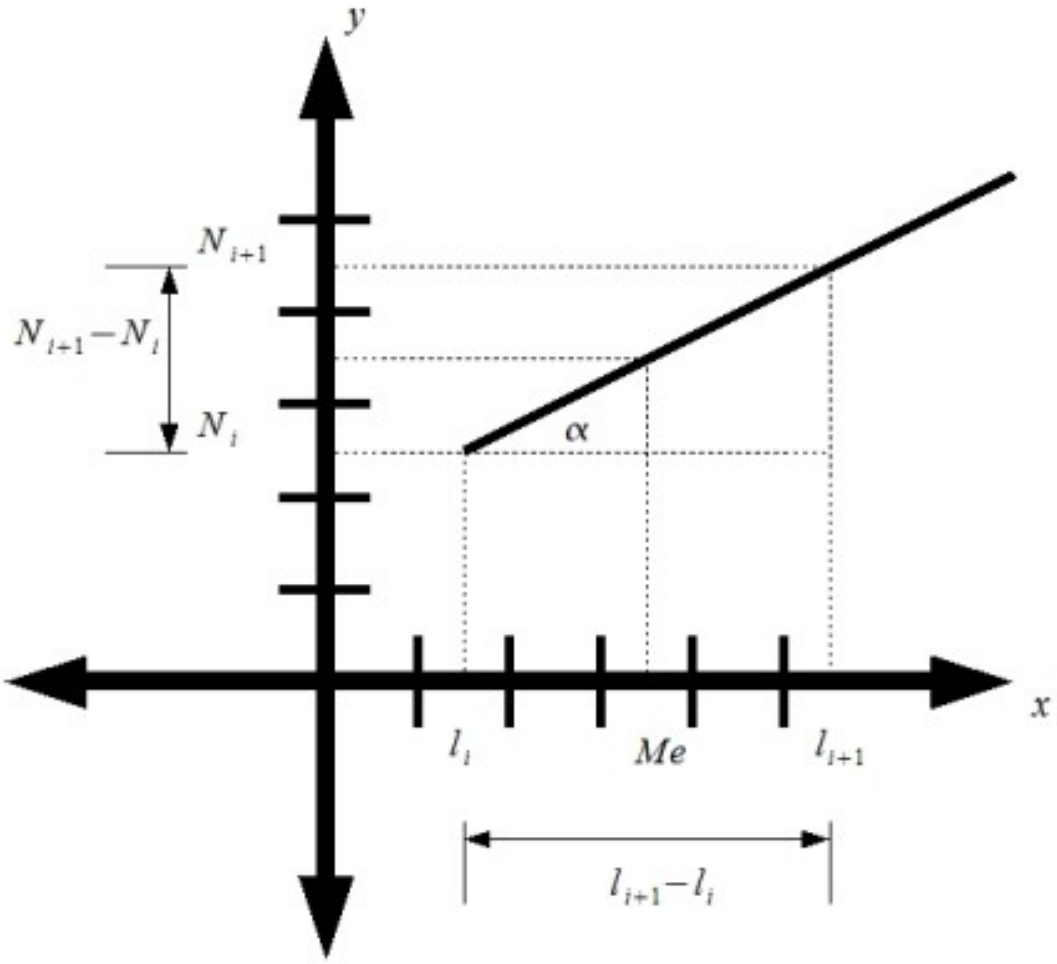


Figura: Proporción lineal para situar la mediana en el intervalo

La idea que hay detrás de este criterio es suponer que los datos de la muestra original se distribuían, antes de agruparlos, de forma uniforme por todo el intervalo. En cualquier caso, esta distancia se puede hallar de varias formas equivalentes:

- Utilizando la semejanza de triángulos entre el triángulo mayor y el menor.
- Utilizando el ángulo α que aparece en el mismo dibujo y la definición de tangente. (En el fondo esta forma es la misma que la anterior, puesto que la tangente de un ángulo se define a partir de un triángulo rectángulo.)

- Con una regla de tres simple directa. Quizá este camino resulte el más fácil, pero en general al aplicar la regla de tres hay que asegurarse que hay una relación lineal (y que «pasa por el origen») entre las cantidades:

$$\begin{array}{ccc} l_{i+1}-l_i & \text{----------} & N_{i+1}-N_i \\ x & \text{----------} & \dfrac{k\cdot N}{100}-N_i \end{array}$$

  Si de esta regla se despeja la variable x, se obtiene la distancia que se le suma al extremo inferior del intervalo seleccionado.

Así es fácil recordar de dónde viene la fórmula y, en caso de que se nos olvide, volver a reconstruirla:

$$P_k=l_i+x=l_i+\frac{\left(\frac{k\cdot N}{100}-N_i\right)\cdot(l_{i+1}-l_i)}{N_{i+1}-N_i}=l_i+\frac{\frac{k\cdot N}{100}-N_i}{N_{i+1}-N_i}(l_{i+1}-l_i)$$

donde, si el intervalo donde está la medida es el *i*-ésimo, l_i es el extremo inferior del intervalo *i*-ésimo y N_i es la frecuencia absoluta acumulada del intervalo *i*-ésimo.

Con saber esta fórmula es suficiente, después basta utilizarla con el valor de *k* adecuado según se quieran calcular percentiles, deciles, cuartiles o la mediana. Por ejemplo: para cuartiles k = 25, 50 o 75, para deciles k = 10, 20... Los percentiles/cuantiles incluyen como casos particulares a esas otras medidas.

Notas:

Si tenemos los datos tabulados (estén sin agrupar o agrupados), tanto las frecuencias acumuladas absolutas como las relativas nos servirían para ver en qué valor o clase está la medida que buscamos. Los dos procedimientos anteriores están descritos para cuando tenemos en la tabla las frecuencias absolutas acumuladas, que van expresando «número de datos» (o posiciones). Se podrían adaptar para cuando se tienen en la tabla las frecuencias relativas acumuladas, utilizando ahora $k/100$, que está entre 0 y 1; no obstante, siempre es posible convertir las frecuencias relativas en frecuencias absolutas. Para la mediana

$$\frac{k\cdot N}{100}=\frac{50\cdot N}{100}=\frac{N}{2}.$$

Por otro lado, las medidas de localización utilizan el orden de los datos en la muestra (cuando la ordenamos de menor a mayor), no directamente de los valores propiamente dichos. Por esto, si los datos de los extremos están muy alejados (muy grandes o muy pequeños), este orden no variará y estas medidas no se verán afectadas por estos datos atípicos. Se dice que son medidas *robustas*. La media, sin embargo, no es robusta: de su expresión, dado que utiliza todos los valores, se ve claramente que si uno de ellos toma un valor muy grande, el resultado final será también grande.

Datos atípicos

En Estadística se considera que un dato es atípico (*outlier*, en inglés) si es muy distinto a los demás y parece generado por otro proceso distinto al de los demás. Como la realidad es complicada y

desconocemos todos los detalles que contribuyen a generar una muestra, no podemos estar seguros de si ese dato se ha debido a que ha sucedido algo de probabilidad pequeña (pero legítimo para el modelo, porque las colas de la mayoría de las distribuciones tienden a cero poco a poco) o es un error que ha aparecido en el proceso de medición, apunte de los datos, transcripción a limpio, etcétera. No hay, por tanto, una regla siempre válida a la hora de decidir qué hacer cuando aparecen datos así. Aunque hay consejos y reglas, en último término la elección se deja a la persona que analiza los datos, que verá en cada caso qué hacer, según las características del experimento, las consecuencias y su experiencia.

En caso de que se quieran quitar los datos atípicos de la muestra, los pasos para quitarlos son:

1. Calcular las medidas que describen las variables. En concreto se necesitan los cuartiles primero y tercero, Q_1 y Q_3.

2. Calcular el rango intercuantílico: $RI = Q_3 - Q_1$.

3. Se suelen considerar *datos atípicos* (no extremos) los que están a la izquierda (son menores) de Q_1 una distancia mayor a 1,5RI, y los que están a la derecha (son mayores) de Q_3 una distancia mayor a esa misma longitud. Es decir, tendremos en cuenta los datos incluidos en el intervalo

$$[Q_1 - 1{,}5 \cdot RI, Q_3 + 1{,}5 \cdot RI]$$

   Los *datos atípicos extremos* serían los que distan, por un lado y por otro, más de tres veces el rango intercuantílico (en vez de una vez y media).

4. Si un sujeto en el que se han medido dos variables (muestra bivariante) presenta algún valor atípico en alguna de las dos, se saca del estudio al sujeto entero.

5. Después de quitar los atípicos de la muestra original, habría que rehacer de nuevo el análisis de los datos, con la nueva muestra, desde el principio.

Puede suceder que después de quitar los primeros datos atípicos, ahora, con los nuevos cuartiles, aparezcan otros que antes no lo eran. Por ejemplo, si en la muestra original había datos cercanos a los cuartiles pero dentro del intervalo de aceptación, es posible que con los nuevos cuartiles queden fuera del nuevo intervalo de aceptación. En general habría que investigar más profundamente qué está pasando, pero a nuestro nivel vamos a convenir que aplicamos el proceso de quitar atípicos sólo una vez. En Ejercicio 1 y Ejercicio 2 se identifican los datos atípicos.

Distribuciones bimodales

Como su propia definición sugiere, la moda no tiene por qué ser única. Una distribución con dos modas se llama *bimodal*. Este tipo de distribuciones puede aparecer cuando una variable estadística ha sido generada por un proceso que realmente proporciona esos dos valores con la misma frecuencia. Esto puede suceder cuando los datos que observamos son el resultado no de un proceso sino de dos procesos subyacentes solapados. Por ejemplo, imaginemos que se mide el tiempo que alumnos y alumnas tardan en recorrer una distancia. Si suponemos que la capacidad física innata es distinta en hombres y mujeres (quien no quiera no está obligado a hacer esta suposición), resulta que habrá, por ejemplo, medidas de centralización, de dispersión y modas para la distribución de los

tiempos de cada uno de estos grupos, de manera que si medimos conjuntamente (sin estratificar por la variable «sexo»), es de esperar que el resultado refleje en sus medidas esa mezcla, en concreto es probable que sea bimodal: una moda provendrá de la distribución de hombre y otra de la de mujeres. Aparte de la capacidad física innata de cada sexo, hay muchos otros factores que hacen que cada distribución tenga una variabilidad, causa de que las distribuciones de cada sexo estén mezcladas y haya mujeres más rápidas que muchos hombres. Dependiendo del grado en que estén mezcladas las dos distribuciones, será más o menos fácil apreciar las distribuciones originales.

Cálculo de la moda

Por definición, la moda de una variable estadística es el valor (o valores) que más frecuentemente toma la variable en la muestra, da igual basarse en la frecuencia absoluta o a la relativa.

- Para una variable sin agrupar es inmediato calcular la moda, porque se puede identificar visualmente observando las frecuencias de los datos.
- Para datos cuantitativos agrupados (los cualitativos agrupados no tienen interés) no tenemos la información de los valores concretos de la variable, sólo las frecuencias de los intervalos. En una situación así la moda se calcula buscando primero el intervalo modal, que es el de mayor frecuencia, y aplicándole después la fórmula siguiente:

$$Mo = l_i + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} \cdot (l_{i+1} - l_i)$$

Esta fórmula se interpreta en el apartado Interpretación de las fórmulas: asimetría, moda.

Intervalos modales en los extremos

Podría suceder que el intervalo modal fuese el primero o el último. Esto no se da frecuentemente, porque lo normal es que la frecuencia vaya disminuyendo para los intervalos laterales (los de los números menores y mayores). Sin embargo, puede darse esta situación, por ejemplo, porque estos intervalos de los lados se hayan tomado demasiado grandes y tiendan a contener muchos datos (habría que revisar la partición).

En cualquier caso, esta situación no ofrece problemas desde el punto de vista técnico. Pensando en el significado de la moda, podemos deducir que lo que habría que hacer sería aplicar la fórmula pensando que hay al lado un intervalo con frecuencia nula; es decir, si el intervalo modal es el primero, pensar como si hubiese otro antes y tomar $n_{i-1} = 0$, mientras que si el intervalo modal es el último, pensar que hubiese otro después y tomar $n_{i+1} = 0$.

Varios intervalos modales

Para saber qué hacer en la situación en que haya varios intervalos modales, pensemos en lo que significa la moda: es, por definición, el valor que más veces ocurre. Por su propia definición puede, por tanto, no ser única; es decir, lo que tiene sentido es aplicar esa misma fórmula a cada uno de los intervalos modales, y de cada uno saldrá un valor de moda. (Nótese que hallar una media de estos valores, aunque puede venir erróneamente sugerido porque es lo que se hace con la mediana, aquí no

tendría sentido, porque podríamos obtener un valor intermedio entre esos dos que en realidad pertenezca a una clase con una frecuencia mucho menor que la de los intervalos modales: podemos imaginarnos esta situación con un dibujo de barras con dos igual de altas y otras intermedias de alturas mucho menores.)

Si esos intervalos modales están separados, es decir, hay otros entre medias, cada uno dará un valor distinto de moda, y todos estos valores son válidos. Es probable que un histograma así provenga de una distribución multimodal (ver el apartado Distribuciones bimodales). Sin embargo, en el caso en que esos intervalos modales estén juntos, podría suceder que la distribución tenga realmente dos máximos muy juntos o uno solo en el centro.

- Si no tenemos acceso a los valores de la variable estadística, sino sólo a la información ya agrupada, no hay manera de distinguir las dos situaciones mencionadas. En este caso habría que recurrir a información externa del problema concreto para intuir en qué caso estamos (la Estadística tiene que utilizar toda la información que tenga a mano, porque la que una muestra puede representar no es mucha).

- Si tenemos la muestra de la variable estadística, podemos dividir esos dos intervalos por la mitad y volver a representar el histograma: si los dos subintervalos centrales tienen mayor probabilidad que los exteriores, tomaremos como moda el valor central que los separa (había un solo máximo), pero si de nuevo los cuatro subintervalos tienen la misma frecuencia aproximadamente, lo más probable es que haya una distribución bimodal subyacente, y lo correcto sería considerar dos modas.

Por tanto, salvo que se disponga de los datos originales sin agrupar, esta fórmula de la moda no es apropiada para el caso de intervalos modales contiguos, puesto que siempre proporciona una única moda central (véase por qué en Interpretación de las fórmulas: asimetría, moda).

Un poco más...

En esta sección se incluyen algunos comentarios de un nivel más alto, que pretenden por un lado facilitar la comprensión de lo visto y de los temas siguientes, y por otro satisfacer la curiosidad de quien quiera ir un poco más allá de lo estrictamente necesario según el temario.

Interpretación de las fórmulas: asimetría, moda

No tenemos costumbre de detenernos a observar las fórmulas, pero hacerlo puede ayudarnos mucho a comprender los conceptos (por qué están definidos así, por qué «funcionan», etcétera).

Asimetría

En el caso de la asimetría, por ejemplo, vamos a ver qué forma tiene su expresión:

$$g_1=\frac{1}{\sigma^3}\frac{1}{N}\sum_{i=1}^{k}(x_i-\bar{x})^3\cdot n_i=\frac{1}{\sigma^3}\sum_{i=1}^{k}(x_i-\bar{x})^3\cdot f_i$$

- Tenemos un sumatorio de cantidades. Está dividido por la desviación típica al cubo, que por su

significado se toma como la raíz cuadrada positiva de la varianza, que también es positiva.

- Cada cantidad del sumatorio es una resta elevada al cubo por un número positivo.
- Cuando elevamos al cubo un número, el resultado es otro número con el mismo signo.

De lo anterior podemos deducir que:

- $g_1<0$ cuando en el sumatorio haya «mucha masa» a la izquierda de la media (dibujo con más datos a la izquierda), es decir, muchas cantidades negativas o pocas pero grandes.
- $g_1>0$ cuando haya «mucha masa» a la derecha de la media (dibujo con más datos a la derecha), es decir, muchos números mayores que la media o pocos pero mucho mayores que la media, lo que hará que su resta, su potencia y el sumatorio sean positivos.
- Por último, $g_1=0$ cuando las cantidades positivas y negativas del sumatorio se cancelen, es decir, cuando los datos tiendan a situarse en igual cantidad a ambos lados de la media (simetría).

Moda

Vamos ahora a interpretar la fórmula de la moda, para el caso de datos (cuantitativos) agrupados en intervalos. Este concepto se define como el valor —o valores, si hay empate— con mayor frecuencia (absoluta o relativa, da igual). En el caso de datos sin agrupar no hay más que buscar entre las frecuencias. En el caso de datos agrupados, tenemos que aplicar la siguiente fórmula:

$$Mo=l_i+\frac{n_i-n_{i-1}}{(n_i-n_{i-1})+(n_i-n_{i+1})}\cdot(l_{i+1}-l_i)$$

donde, si el intervalo modal es el *i*-ésimo, l_i es el extremo inferior del intervalo modal, n_i es la frecuencia absoluta del intervalo modal y n_{i-1} e n_{i+1} son, respectivamente, las frecuencias absolutas de los intervalos anterior y posterior al modal.

- Si observamos la fórmula vemos que la moda se calcula tomando como referencia el extremo inferior del intervalo modal (el *i*-ésimo), al que se le suma una cantidad.
- Esa cantidad que se le suma es la amplitud del intervalo multiplicada por un número.
- Ese número es un cociente; más concretamente, una proporción. Si el intervalo modal es el *i*-ésimo, se cumple que $n_i>n_{i-1}$ y $n_i>n_{i+1}$, es decir, la frecuencia del intervalo modal es mayor que las de los intervalos contiguos. Así, se tiene que $n_i-n_{i-1}>0$ y $n_i-n_{i+1}>0$. Es decir que el cociente por el que se multiplica a la amplitud es positivo y menor que uno, porque podemos ver que el denominador consta de la misma cantidad que el numerador más otra cantidad que es positiva.
- Si cogemos el extremo inferior y sumamos una cantidad menor que la amplitud, no puede suceder que el resultado sea una cantidad que se salga del intervalo modal. Pero si el cociente anterior es cero o uno, el valor que se obtiene es el de los extremos del intervalo modal; esto sucede cuando hay dos intervalos modales contiguos, léase lo que se dice en el apartado Cálculo de la moda sobre esto.

Respecto a la interpretación, no es tan clara como la de la fórmula de los percentiles, pero

podemos ver que la moda es el valor que dista del extremo inferior del intervalo una longitud que es una proporción de la amplitud; y esta proporción es la que hay entre lo que aumenta la frecuencia del intervalo anterior al modal con relación a lo que ha aumentado del anterior al modal más lo que aumenta respecto al siguiente. Más sencillo que con palabras,

$$\begin{array}{ccc} (n_i - n_{i-1}) + (n_i - n_{i+1}) & \text{----------} & 1 \\ (n_i - n_{i-1}) & \text{----------} & x \end{array}$$

Y entonces

$$Mo = l_i + x \cdot (l_{i+1} - l_i).$$

Vemos que cuando hay dos intervalos modales seguidos, al aplicar la fórmula a cada uno sucede que $x=1$ para el intervalo de la izquierda y $x=0$ para el de la derecha, con lo que en ambos casos se obtiene el punto que divide los dos intervalos. Esta fórmula no es adecuada para discernir, en este caso de intervalos modales contiguos, entre una distribución unimodal y una bimodal (véase lo dicho en Cálculo de la moda).

MEDA

Un ejemplo de una medida de variabilidad en que no se mide la distancia de los datos respecto a la media (como medida de centralización), sino respecto a la mediana, es la MEDA. En esta medida se calcula la distancia de cada dato a la mediana de la muestra, se quita el signo en todas las cantidades anteriores tomando valores absolutos y finalmente se toma la mediana de estas cantidades.

El inconveniente de la MEDA es que es mucho más costosa de calcular, pero la ventaja es que no depende de datos atípicos (muy distintos a la mayoría de los demás), porque la mediana tampoco depende de estos datos, mientras que la media sí lo hace. Obsérvese que la mediana depende del orden de los datos, no directamente de los propios valores de los datos.

Coeficiente de apertura

El *coeficiente de apertura* se define como el cociente entre el mayor valor de la muestra y el menor.

$$CA = \frac{máx\{x_i\}}{mín\{x_i\}}.$$

Por otro lado el *rango* se define como:

$$R = máx\{x_i\} - mín\{x_i\}.$$

Ambos conceptos son medidas de dispersión. Podemos apreciar que el rango tiene las mismas unidades que la variable estadística, porque si cada dato está dado en m^2, por ejemplo, su diferencia también estará en esa unidad. Sin embargo, el coeficiente de apertura no tiene unidades, porque el cociente de dos cantidades dadas en m^2 es un número sin unidades. Para aclararnos con las unidades,

se puede pensar que son números:

$$3\text{m}^2+6\text{m}^2=(3+6)m^2=9\text{m}^2, \qquad 3\text{m}\cdot 6\text{m}^2=(3\cdot 6)m\cdot m^2=18\text{m}^3 \qquad \text{y} \qquad \frac{3\text{m}^2}{6\text{m}^2}=\frac{3\cancel{m^2}}{6\cancel{m^2}}=\frac{3}{6}=\frac{1}{2}$$

En cuanto a la utilidad de estas dos medidas, no se puede decir que una sea más útil que otra, dependerá del caso concreto, de lo que represente la variable estadística. Hay casos en los que utilizar una escala de diferencias es mejor y otros en que lo es una escala de cocientes. Salvo que una tiene unidades y otra no (no cambia si se transforman los datos a otra unidad), la diferencia principal estará en la interpretación:

- Que $R=15$ significa que el máximo de los datos se sitúa quince unidades de medida más a la derecha que el mínimo, el resto de los datos está entre estas cantidades y la variable se dispersa o distribuye en ese rango. En este caso hay un patrón de unidad de medida.

- Que $CV=15$ significa que el máximo de los datos es quince veces mayor que el mínimo. Aquí el mínimo está como haciendo de patrón de unidad con el que se compara el máximo.

Algunos ejercicios resueltos

Ejercicio 1

En una pequeña empresa con 60 empleados, 25 son personal de fábrica y están cobrando unos sueldos semanales (en euros) en función de su antigüedad de:

300 250 241 360 295 316 296 311 318 345 310 294 240

266 299 298 300 318 317 317 314 293 315 316 340

El resto del personal trabaja en oficina y tiene un sueldo medio de 336,57 euros por semana, y una desviación típica de 50,00 euros.

a) Calcular la media y la desviación típica del sueldo por semana del personal de fábrica.

b) Representar el diagrama de tallo y hojas del personal de fábrica.

c) Representar el diagrama de caja del personal de fábrica. ¿Existen datos atípicos?

d) Calcular el sueldo medio por semana de todo el personal.

e) Supongamos que se dobla el sueldo a todos los empleados y a continuación, por la excesiva subida, se decide quitarles 275,00 euros por semana. Calcular las nuevas medias y desviaciones típicas del sueldo semanal correspondiente al personal de fábrica y personal de oficina.

Tenemos la muestra aleatoria simple $(x_{1,}\cdots,x_{25},x_{26},\cdots,x_{60})$, que podemos dividir en dos submuestras:

$(x_1,\cdots,x_{25})$ $\equiv$ Personal de fábrica

$\bar{x}_1=¿$ euros/semana

$s_1=\sqrt{s_1^2}=¿$ euros/semana

$(x_{26},\cdots,x_{60})=¿$ $\equiv$ Personal de oficina

$\bar{x}_2=336{,}57$ euros/semana

$s_2=\sqrt{s_2^2}=50{,}00$ euros/semana

Vemos que de la primera submuestra conocemos los datos originales, mientras que de la segunda sólo conocemos las medidas que nos dan; de hecho, de esta segunda submuestra es imposible recuperar los valores de los datos originales.

a) $\bar{x}_1=\dfrac{x_1+\cdots+x_{25}}{25}=\dfrac{300+\cdots+340}{25}=302{,}76$ euros/semana

$$s_1^2=\frac{(x_1-\bar{x})^2+\cdots+(x_{25}-\bar{x})^2}{25}=\frac{(300-302{,}76)^2+\cdots+(340-302{,}76)^2}{25}=29{,}07 \text{ euros}^2/\text{semana}^2$$

$s_1=\sqrt{s_1^2}=\sqrt{29{,}07}=5{,}39$ euros/semana

Como vemos, las unidades de la varianza son el cuadrado de las de los datos.

b) Diagrama de tallos y hojas (unit = 1,0 1|2 represents 12,0)

```
     LO|240,0 241,0 250,0

  3   24|
  3   25|
  4   26|6
  4   27|
  4   28|
 10   29|345689
 12   30|00
(10)  31|0145667788
  3   32|
  3   33|
```

3 34|05

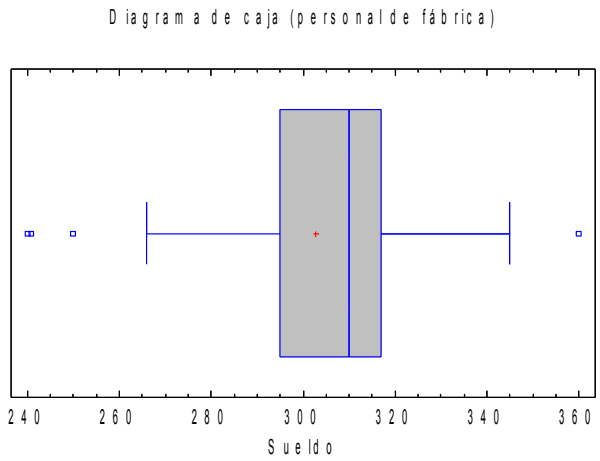
HI|360,0

c) Como tenemos el diagrama de tallos y hojas, vamos a aprovecharlo para obtener la muestra ordenada de menor a mayor (si no tuviésemos el diagrama, la podemos ordenar directamente de la muestra original, sin necesidad de hacer el diagrama). De hecho, una de las ventajas de este tipo de diagrama es que permite recuperar los valores de la muestra, cosa que no permiten hacer otros gráficos:

(240, 241, 250, 266, 293, 294, **295**, 296, 298, 299, 300, 300, **310**, 311, 314, 315, 316, 316, 317, **317**, 318, 318, 340, 345, 360)

La información que necesitamos para representar el diagrama es:

- Primer cuartil: $Q_1 = 295$
- Mediana o segundo cuartil: $M = 310$
- Tercer cuartil: $Q_3 = 317$
- Rango intercuantílico: $Q_3 - Q_1 = 317 – 295 = 22$
- Límite admisible inferior: $LI = Q_1 - 1{,}5RI = 295 – 33 = 262$
- Límite admisible superior: $LS = Q_3 + 1{,}5RI = 317 + 33 = 350$
- Datos atípicos: Por debajo de 262 están los datos 240, 241 y 250. Por encima de 350 está el valor 360



Vemos que el programa informático utilizado para hacer este diagrama no dibuja los límites admisibles. En cualquier caso, nosotros tenemos que calcularlos para identificar los datos atípicos.

d) Veamos una propiedad que siempre es cierta:

$$\bar{x}=\frac{1}{n}(x_1+\cdots+x_n)=\frac{1}{n}(x_1+\cdots+x_k)+\frac{1}{n}(x_{k+1}+\cdots+x_n) \quad =$$

$$= \frac{k}{n}\frac{1}{k}(x_1+\cdots+x_k)+\frac{n-k}{n}\frac{1}{n-k}(x_{k+1}+\cdots+x_n)=\frac{k}{n}\bar{x}_1+\frac{n-k}{n}\bar{x}_2$$

En nuestro caso:

$$\bar{x}=\frac{1}{60}(x_1+\cdots+x_{60})=\frac{1}{60}(x_1+\cdots+x_{25})+\frac{1}{60}(x_{26}+\cdots+x_{60}) \quad =$$

$$= \frac{25}{60}\frac{1}{25}(x_1+\cdots+x_{25})+\frac{35}{60}\frac{1}{35}(x_{26}+\cdots+x_{60})=\frac{25}{60}\bar{x}_1+\frac{35}{60}\bar{x}_2 \quad =$$

$$= \frac{25}{60}302{,}76+\frac{35}{60}336{,}57=126{,}15+196{,}33=322{,}48 \quad \text{euros/semana}$$

e) Veamos por separado el efecto de algunas transformaciones de los datos:

Transformación 1

$$(x_1,\cdots,x_n) \quad \rightarrow \quad (y_1,\cdots,y_n)=(\alpha\cdot x_1,\cdots,\alpha\cdot x_n)$$

Multiplicar o dividir por un número

Esta operación, que se puede llamar *escalamiento* o, más técnicamente, *homotecia* (http://es.wikipedia.org/wiki/Homotecia), equivale a cambiar la escala que mide las distancias entre los datos. Por tanto, intuitivamente:

1. La media, por ser la suma de todos los valores, queda transformada de la misma manera.
2. Las medidas de variabilidad (hablamos aquí de la varianza y la desviación típica) cambian, puesto que cambian las distancias relativas de cada dato a la media. En concreto, la varianza queda multiplicada o dividida por la misma constante al cuadrado (por haber un cuadrado en su definición), mientras que la desviación, por tener después del cuadrado una raíz cuadrada, se ve afectada por la misma constante que los datos sin elevar al cuadrado.

Lo vemos analíticamente:

Media:

$$\bar{y}=\frac{1}{n}(y_1+\cdots+y_n)=\frac{1}{n}(\alpha\cdot x_1+\cdots+\alpha\cdot x_n)=\alpha\cdot\frac{1}{n}(x_1+\cdots+x_n)=\alpha\cdot\bar{x}$$

Varianza:

$$s_y^2=\frac{1}{n}[(y_1-\bar{y})^2+\cdots+(y_n-\bar{y})^2]=\frac{1}{n}[(\alpha x_1-\alpha\bar{x})^2+\cdots+(\alpha x_n-\alpha\bar{x})^2] \quad =$$

$$= \frac{1}{n}[\alpha^2(x_1-\bar{x})^2+\cdots+\alpha^2(x_n-\bar{x})^2]=\alpha^2\cdot\frac{1}{n}[(x_1-\bar{x})^2+\cdots+\alpha^2(x_n-\bar{x})^2]=\alpha^2\cdot s_x^2$$

Desviación típica:

$$s_y=\sqrt{s_y^2}=\sqrt{\alpha^2\cdot s_x^2}=\alpha\cdot s_x$$

Transformación 2

$$(x_1,\cdots,x_n) \quad \rightarrow \quad (y_1,\cdots,y_n)=(x_1-\beta,\cdots,x_n-\beta)$$

Sumar o restar un mismo valor

Esta operación se llama *traslación*, y, si nos imaginamos los valores representados en una recta horizontal, equivale a moverlos todos para un lado u otro, dependiendo de si se suma o resta un valor. Intuitivamente:

1. La media, que es el «centro de gravedad» de los puntos (podemos imaginarlos como masas, con peso) se ve afectada por la misma traslación que cada dato de la muestra.
2. Las medidas de variabilidad (varianza y desviación típica), que se forman a partir de distancias relativas a la media, no cambian cuando trasladamos los datos, porque esas distancias relativas se conservan.

Lo demostramos analíticamente:

Media:

$$\bar{y}=\frac{1}{n}(y_1+\cdots+y_n)=\frac{1}{n}[(x_1-\beta)+\cdots+(x_n-\beta)]=\frac{1}{n}[(x_1+\cdots+x_n)-(\beta+\cdots+\beta)] =$$

$$= \frac{1}{n}[(x_1+\cdots+x_n)-(\beta+\cdots+\beta)]=\frac{1}{n}(x_1+\cdots+x_n)-\frac{1}{n}(\beta+\cdots+\beta)=\bar{x}-\beta$$

Varianza:

$$s_y^2=\frac{1}{n}[(y_1-\bar{y})^2+\cdots+(y_n-\bar{y})^2]=\frac{1}{n}[(x_1-\beta-\bar{x}+\beta)^2+\cdots+(x_n-\beta-\bar{x}+\beta)^2] =$$

$$= \frac{1}{n}[(x_1-\cancel{\beta}-\bar{x}+\cancel{\beta})^2+\cdots+(x_n-\cancel{\beta}-\bar{x}+\cancel{\beta})^2]=\frac{1}{n}[(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2]=s_x^2$$

Desviación típica:

$$s_y=\sqrt{s_y^2}=\sqrt{s_x^2}=s_x$$

Transformación 3

Para la transformación

$$(x_1, \cdots, x_n) \quad \rightarrow \quad (y_1, \cdots, y_n) = (a \cdot x_1 - b, \cdots, a \cdot x_n - b)$$

se puede pensar que primero se aplica la transformación 1 y después la transformación 2 (esta operación de aplicar una transformación tras otra se llama *componer* en Matemáticas). Los cambios en las medidas son:

$$\bar{x} \quad \rightarrow \quad a \cdot \bar{x} \quad \rightarrow \quad a \cdot \bar{x} - b$$

$$s_x^2 \quad \rightarrow \quad a^2 \cdot s_x^2 \quad \rightarrow \quad a^2 \cdot s_x^2$$

$$s_x \quad \rightarrow \quad a \cdot s_x \quad \rightarrow \quad a \cdot s_x$$

Entonces:

$$\bar{y} = a \cdot \bar{x} - b$$

$$s_y^2 = a^2 \cdot s_x^2$$

$$s_y = a \cdot s_x$$

En nuestro caso, como $a = 2$ y $b = 275$, se tiene que:

$\bar{y}_1 = 2 \cdot \bar{x}_1 - 275 = 2 \cdot 302{,}76 - 275 = 330{,}52$ euros/semana

$s_{y,1}^2 = 2^2 \cdot s_1^2 = 2^2 \cdot 29{,}07 = 116{,}28$ euros^2/semana^2

$s_{y,1} = \sqrt{116{,}28} = 10{,}78$ euros/semana

y

$\bar{y}_2 = 2 \cdot \bar{x}_2 - 275 = 2 \cdot 336{,}57 - 275 = 398{,}14$ euros/semana

$s_{y,2}^2 = 2^2 \cdot s_2^2 = 2^2 \cdot 50^2 = 10000{,}00$ euros^2/semana^2

$s_{y,2} = 2 \cdot s_2 = 2 \cdot 50 = 100{,}00$ euros/semana

Ejercicio 2

El año pasado, las duraciones de los cortes publicitarios en las películas emitidas por la televisión el lunes a las 10 de la noche, tomaron valores alrededor de un valor medio de 9 minutos. Para comprobar la actualidad de este valor medio, se midieron durante un mes las duraciones de varios cortes publicitarios. Los datos obtenidos se presentan en la tabla siguiente:

9,2 9,8 9,3 8,3 12,4 8,2 15,3 8,6 8,8 13,5 8,6 7,9

a) Calcular la media y la mediana de este conjunto de datos.

b) Deducir de los valores de estos dos parámetros de localización el aspecto esquemático de la distribución de esas duraciones.

c) Representar los datos mediante un diagrama de caja.

d) Se decide omitir los datos atípicos del conjunto de datos iniciales. Calcular los parámetros obtenidos en a) para este nuevo conjunto de datos.

a) Si la muestra es $(x_{1,}\cdots,x_{12})$, la media es

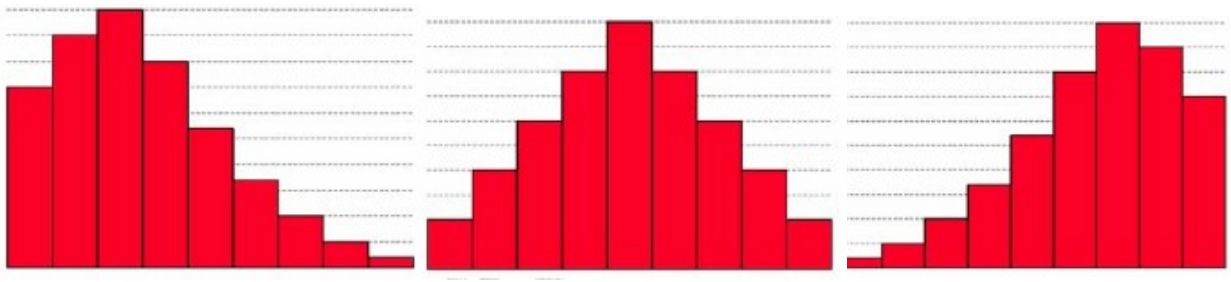
$$\bar{x}=\frac{x_1+\cdots+x_{12}}{12}=10{,}0 \quad \text{minutos}$$

Ahora, si la muestra ordenada de menor a mayor es $(x_{(1)},\cdots,x_{(12)})$, la mediana es

$$M=\frac{x_{(6)}+x_{(7)}}{2}=9{,}0 \quad \text{(dado que en la muestra hay un número par de datos).}$$

Por sus expresiones, vemos que tanto la media como la mediana pueden ser distintas a todos los datos de la muestra, no tienen por qué coincidir con ninguno de ellos.

b) Como también vemos, la media y la mediana no coinciden; es decir, la variable estadística no se distribuye simétricamente. La media es mayor que la mediana, lo que es característico de las distribuciones asimétricas a la derecha. Veamos los siguientes histogramas teóricos:



De izquierda a derecha, estos histogramas corresponden a muestras de variables cuya distribución es, respectivamente, asimétrica a la derecha, simétrica y asimétrica a la izquierda.

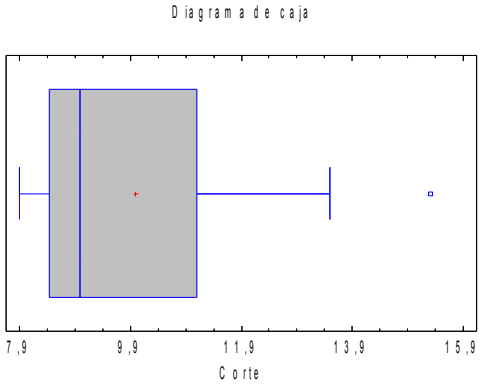
c) En este caso, como hay 12 datos, el 25% supone 3 datos, así que, si la muestra ordenada es:

7,9 8,2 **8,3** **8,6** 8,6 **8,8** **9,2** 9,3 **9,8** **12,4** 13,5 15,3

- Primer cuartil: $Q_1 = x_{(4)} = 8{,}6$. Hay quien aplica el criterio de tomar la semisuma de los valores contiguos a la posición de interés, este caso sería:

$$Q_1 = \frac{x_{(3)} + x_{(4)}}{2} = \frac{8{,}3 + 8{,}6}{2} = 8{,}5.$$

  Ambos criterios son válidos, puesto que lo importante no es tanto el valor concreto como la información que da sobre cómo se distribuyen los datos de la muestra.
- Mediana o segundo cuartil: M = (8,8 + 9,2)/2 = 9,0
- Tercer cuartil: Q_3 = (9,8 – 12,4)/2 = 11,1
- Rango intercuantílico: Q_3 - Q_1 = 11,1 – 8,45 = 2,7
- Límite admisible inferior: LI = Q_1 - 1,5RI = 8,45 – 3,97 = 4,5
- Límite admisible superior: LS = Q3 + 1,5RI = 11,1 + 3,97 = 15,1
- Datos atípicos: Por debajo de 4,5 no hay datos. Por encima de 15,1 está el valor 15,3, que es, por tanto, el único dato atípico.



d) Al omitir el dato atípico, la muestra quedaría:

7,9 8,2 8,3 8,6 8,6 **8,8** 9,2 9,3 9,8 12,4 13,5

Para esta nueva muestra se tiene que:

$$\bar{x} = \frac{x_{(1)} + \cdots + x_{(11)}}{11} = 9{,}5 \quad \text{minutos}$$

$$M = x_{(6)} = 8{,}8$$

Vemos que la mediana ha cambiado menos que la media; esto se debe a que es *robusta*, mientras que la media no lo es. Por otro lado, es importante hacer notar que la mediana ha cambiado de valor porque hemos omitido un dato, pero no cambiaría de valor si hubiésemos sustituido ese valor extremo por otro aún más extremo, por ejemplo, puesto

que la mediana se elige por las posiciones de la muestra ordenada, no por los valores de los datos.

Dos variables

Características de los datos

Dos muestras univariantes y una muestra bivariante

Dada una población, conceptualmente es importante distinguir entre las dos situaciones siguientes:

1) Considerar una muestra de elementos de la población y medir en ellos la variable A. Luego considerar otra muestra distinta de elementos de la población y medir en ellos la variable B.
2) Considerar una muestra de la población y medir en sus elementos las dos variables A y B.

En el primer caso se obtendrán dos *muestras univariantes*, $(a_1,...a_n)$ y $(b_1,...b_m)$, con las que en algunos casos se pueden construir a posteriori pares de valores (cuando tenga sentido: por ejemplo, cuando exista un conjunto de clases y las variables a_i y b_i se midan en sujetos distintos pero de la misma clase), mientras que en el segundo caso se obtendrá directamente una *muestra bivariante*, $(a_1, b_1),...,$ (a_n, b_n).

A la hora de estudiar la relación entre las dos variables, si las muestras de la situación 1 no son muy distintas entre sí, cosa que se puede controlar sólo hasta cierto punto, sus resultados diferirán poco de los que se puedan deducir de la situación 2. Sin embargo, en general se corre el riesgo de confundir diferencias entre las muestras con diferencias entre las variables; es algo que hay que tener en cuenta a la hora de interpretar los análisis de los datos. Por tanto, siempre que sea posible, la manera más adecuada de estudiar la relación entre dos variables es la segunda situación descrita.

Vuelta a la estadística de una variable

Quien tiene cierta cantidad de información puede decidir no utilizarla toda. Con esta obviedad sólo se quiere introducir una relación trivial entre la estadística de dos variables y la de una variable. Todo lo que se ve en los temas de Estadística descriptiva de una variable es aplicable a las distintas variables estadísticas que «forman parte de» o «se pueden construir a partir de» dos variables estadísticas, por ejemplo:

(a) Se puede tener en cuenta sólo la primera variable

(b) Se puede tener en cuenta sólo la segunda variable

(c) Se puede tener en cuenta la primera variable sólo en los casos en que la segunda variable toma cierto valor

(d) Se puede tener en cuenta la segunda variable sólo en los casos en que la primera variable toma cierto valor

A cada una de estas variables estadísticas, como variables en sí mismas, les es aplicable todo lo que la Estadística descriptiva de una variable: características de las variables, agrupamiento en clases, distintos tipos de frecuencias, expresión en tablas, representación en gráficos, cálculo de medidas numéricas, etcétera.

La información correspondiente a dos variables puede venir dada como un conjunto de pares de valores, tal y como se han medido para cada sujeto de la muestra, o ya expresadas en una *tabla de doble entrada* (una entrada o dirección para cada una). En el primer caso, los ejemplos (a) y (b) consistirían en tener en cuenta, en todos los pares, sólo el valor de una misma posición, mientras los ejemplos (c) y (d) consistirían en observar sólo los valores de una posición para los pares en que la otra tome un cierto valor. En el segundo caso, de los ejemplos anteriores el (a) y el (b) son las *distribuciones marginales*, mientras que (c) y (d) son las diferentes *distribuciones condicionadas*.

Por todo lo anterior, no se insistirá más en este documento en que la Estadística descriptiva de una variable es plenamente aplicable a cada una de estas variables estadísticas individuales. Entonces, para que la Estadística descriptiva de dos variables tenga más valor que el de la simple aportación de dos estadísticas de una variable, tiene que tener como principal objetivo estudiar la relación o interacción entre las dos variables.

Un poco más...

En esta sección se incluyen algunos comentarios de un nivel más alto, que pretenden por un lado facilitar la comprensión de lo visto y de los temas siguientes, y por otro satisfacer la curiosidad de quien quiera ir un poco más allá de lo estrictamente necesario según el temario.

Datos multivariantes y datos multidimensionales

No siempre se tiene el debido cuidado a la hora de utilizar los términos «multivariante» y «multidimensional». La diferencia (que he leído en buenos autores) está en que:

A) Medir varias variables da lugar a *datos multivariantes*. Por ejemplo, se pueden medir la altura y el peso de una persona: (A,P).

B) Medir una cantidad que depende de varias variables da lugar a *datos multidimensionales*. Por ejemplo, medir la temperatura depende de la posición y el tiempo: $T(x,t)$.

Quizá a veces la confusión viene de que se puede pensar que T es multidimensional y depende del vector multivariante (x,t). Por otro lado, para un vector aleatorio o variable aleatoria bivariante se considera la función de densidad conjunta, $f(x,y)$, que es bidimensional.

Obviamente, también es posible combinar los dos tipos de datos anteriores para obtener datos multivariantes y multidimensionales, por ejemplo, medir la temperatura y la presión: $(T(x,t),P(x,t))$.

Por otro lado, para seguir aclarando ideas, en vez de una sola medición se pueden hacer varias, ordenadas siguiendo el orden de algún índice: si este índice es el tiempo, se habla de *procesos temporales*, mientras que si el índice es una posición espacial, se habla de *procesos espaciales*. Si el índice es el tiempo y sólo se lleva a cabo una medición, pero se entiende que en varios sujetos (una

muestra), se suele hablar de una medida *simultánea* o *transversal*; si se hacen varias mediciones ordenadas por el índice (tiempo u otra variable), se habla a veces de medida *longitudinal*.

En cualquier caso, con este pequeño apartado sólo se quería por un lado defender el enriquecimiento de pensamiento (y del idioma) que significa distinguir los términos «multivariante» y «multidimensional», y por otro justificar por qué en este documento se trabaja con *datos bivariantes* y no con datos bidimensionales.

Vectores aleatorios

De la misma manera que a las variables estadísticas univariantes se les podían asignar, como modelo, las variables aleatorias, a las variables estadísticas bivariantes se les puede asignar una *variable aleatoria bivariante*, que no será más que un vector con dos variables aleatorias con cierta relación entre ellas (si no hay relación, la situación no tiene mucho interés en la estadística con la que estamos trabajando ahora).

Construcción de tablas

Distribuciones conjunta, marginales y condicionales

Cuando se estudia una variable estadística, se suele utilizar una tabla para recoger la información de la muestra. En esta tabla, en una dirección se expresan los valores de la variable (o de las clases, en el caso de que se hubiese agrupado la muestra), y dentro, paralelamente, se incluyen las frecuencias absolutas, relativas o sus acumuladas.

En este capítulo se estudian dos variables a la vez, por lo que la información se va a recoger también en tablas, con la diferencia de que ahora necesitamos las dos direcciones —vertical y horizontal— para recoger en cada una los posibles valores de una variable (o clases, si se ha agrupado). La información de las celdas será también las frecuencias, en este caso *conjuntas*. Sin embargo, en el caso multivariante hay otras frecuencias además de las absolutas, relativas y sus acumuladas: las *frecuencias marginales* y las *frecuencias condicionales*.

Distribución conjunta

En el estudio conjunto de dos variables estadísticas hay un interés especial en estudiar la relación entre las variables, pues en otro caso se estarían haciendo, como se ha mencionado, dos estudios independientes, cada uno de ellos para una sola variable. La *distribución conjunta*, como su nombre indica, es la distribución que explica conjuntamente el comportamiento de ambas variables a la vez. Es decir, cada celda de la tabla inicial nos está informando de la frecuencia n_{ij} con que ha sucedido a la vez que la variable X ha tomado un valor de su clase i-ésima y la variable Y lo ha tomado de su clase j-ésima (lo que también se puede interpretar como la intersección de dos sucesos). La representación gráfica equivalente al histograma de una variable, que se representaba en el plano, es ahora un histograma tridimensional, que se representa en el espacio de tres dimensiones (la altura de las barras depende de dos dimensiones: es una función bidimensional). Pero un gráfico más

sencillo que éste es el *diagrama de dispersión* o *nube de puntos*, del que se habla en otro apartado. Volviendo a los histogramas, con la siguiente figura se quiere mostrar la diferencia entre estudiar las dos variables por separado, hacer sus gráficos y solaparlos, y estudiar conjuntamente las variables para representar la relación en un histograma tridimensional.

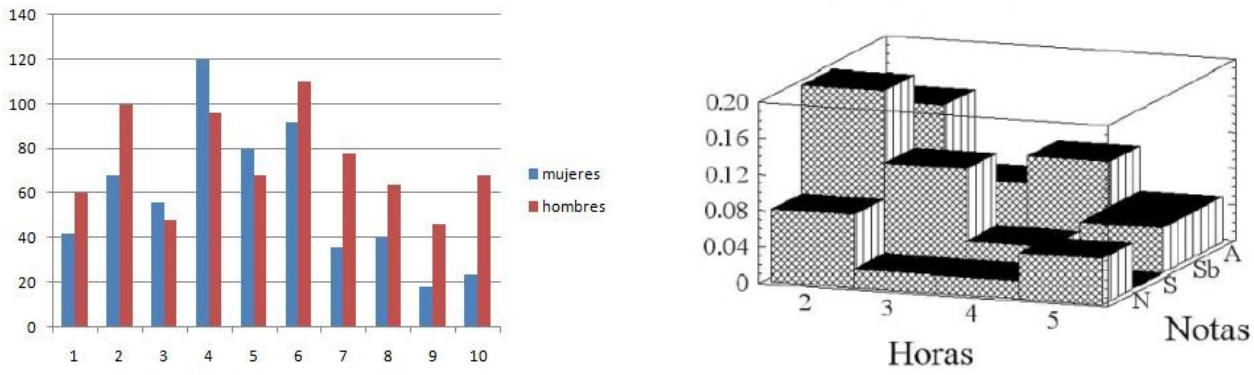


Figura: Dos histogramas independientes y un histograma conjunto en tres dimensiones

Distribuciones marginales

La *distribución marginal* de una variable, como su nombre también indica, se representa en el margen de la tabla. Habrá, por tanto, una distribución marginal para cada una de las dos variables en cada uno de los márgenes. Estas distribuciones se calculan sumando todas las frecuencias de cada columna y de cada fila. Esta suma significa estadísticamente que agrupamos todas las frecuencias en una, es decir, que nos da igual qué valor haya tomado la otra variable, no tenemos en cuenta su agrupamiento en clases. Las frecuencias que obtenemos en cada margen nos informan de cómo se ha agrupado la muestra en las clases de esa variable, sin atender a cómo lo hayan hecho en las clases de la otra variable. Esta información debe ser cercana (salvo pequeñas diferencias) a la que se obtendría de dos estudios independientes que estuviesen dirigidos cada uno a estudiar sólo una variable.

Distribuciones condicionales

Por último, las *distribuciones condicionales*, como su nombre indica, son las que nos informan de cómo se distribuye una variable cuando la otra variable pertenece a una clase determinada; es decir, informa de la distribución de una variable condicionada a un determinado valor de la otra. Para calcular una distribución condicionada, como la idea es estudiar cómo se distribuyen los elementos en las clases de una variable fijada la otra, basta con tomar los valores de la columna o fila adecuada; pero, como sólo una parte de la muestra total ha tomado el valor fijo de la variable condicionada, tomando esa columna o fila estamos considerando cómo se distribuye esta submuestra, y necesitamos normalizar adecuadamente (para que esa submuestra haga de muestra). Por tanto, los cálculos consisten en dividir las frecuencias conjuntas por la frecuencia marginal.

Gráficamente, a partir del histograma tridimensional, es fácil imaginarse lo que significan las distribuciones condicionales: tomar la sección de la fila o columna de barras correspondiente y, al normalizar, cambiarla proporcionalmente para que el área total de las barras sume uno.

Las distribuciones marginales y condicionales se suelen definir a partir de las frecuencias relativas, no de las absolutas. Esto se suele hacer así por la interpretación posterior de las frecuencias relativas y su analogía con las probabilidades (aunque quizá el concepto de normalización que se ha mencionado antes se entiende mejor si hablamos de las absolutas, porque representan «número de

datos»). Para ver que los resultados (por ejemplo, en el caso de las frecuencias condicionadas) son los mismos si se parte de la tabla de las frecuencias relativas que si se parte de la de frecuencias absolutas:

$$\frac{f_{ij}}{f_{\cdot j}}=\frac{\frac{n_{ij}}{N}}{\frac{n_{\cdot j}}{N}}=\frac{n_{ij}\cdot N}{N\cdot n_{\cdot j}}=\frac{n_{ij}\cdot \cancel{N}}{\cancel{N}\cdot n_{\cdot j}}=\frac{n_{ij}}{n_{\cdot j}}$$

La información de cada distribución condicionada es la misma (salvo pequeñas diferencias) que se obtendría de un estudio cuya población fuese directamente los elementos para los que la variable por la que se condiciona tomase un determinado valor. Por otra parte, se menciona una vez más que para cada distribución marginal o condicionada de las anteriores, por ser ella misma una distribución, puede estudiarse la variable estadística que le corresponde como se hace para una única variable.

Representación en gráficos

Diagrama de dispersión

Cuando tenemos dos variables, podemos estudiar su relación analítica y gráficamente. Esto segundo se puede hacer mediante un histograma tridimensional (ver un ejemplo en Distribuciones conjunta, marginales y condicionales); pero es difícil de representar, por lo que se utiliza mucho el *diagrama de dispersión* o *nube de puntos*. Como queremos un gráfico que nos informe de esa posible relación entre las variables, tenemos que representarlas juntas. Podríamos pensar en:

(a) Coger la muestra A y representar los puntos

$$(1,a_1),(2,a_2),\ldots,(N,a_N).$$

Coger la muestra B y hacer lo mismo, con lo que aparecerían en el mismo gráfico los puntos

$$(1,b_1),(2,b_2),\ldots,(N,b_N).$$

Y de este gráfico podríamos intentar intuir la relación entre las dos variables, es decir, qué valor tiene a tomar la una en función de la otro. Esta tarea es normalmente difícil, como muestra el siguiente ejemplo. A partir de una tabla se representan dos variables que parecen distribuirse alrededor de sendas rectas que divergen (esto sucede, por ejemplo al representar las variables AGRICU y ESPM del fichero de datos «paises.txt» adjunto al libro *Introducción a la Estadística para las Ciencias Sociales*, de Daniel Peña y Juan Romo). ¿Qué significa que esas dos muestras se sitúen en torno a rectas que divergen? El que se trate de rectas significa que los datos en la tabla estaban colocados de manera que al descender por sus filas ambas variables van variando linealmente (y, por tanto, por variar linealmente con respecto a la posición en la tabla, varían linealmente entre sí. Escríbase esto analíticamente con las ecuaciones de las rectas...). El que diverjan significa que a medida que una variable crece la otra decrece, es decir, que la correlación (lineal) que hay es negativa. Vemos qué indirectamente aparece en el gráfico esta correlación negativa. Por supuesto, también habría que tener en cuenta si los datos se habían colocado en la tabla con algún criterio.

(b) Otra opción es representar los puntos de la forma

$$(a_1,b_1),(a_2,b_2),...,(a_N,b_N).$$

En las primeras posiciones se ponen los valores de la variable que queramos poner en el eje X, y en las segundas los de la del eje Y. Entonces cada punto nos dice cuánto valen las dos variables para su sujeto de la muestra. Esto es lo que es más propiamente dicho el *diagrama de dispersión*: representar los valores de las dos variables que se han medido para cada sujeto. A partir de él, hay casos en que se ve a simple vista no sólo si hay relación entre las dos variables, sino si además esa relación es lineal o no.

(c) Cosa distinta es que tuviésemos dos muestras univariantes distintas, A y B, no una muestra bivariante (sobre esta diferencia se habla en el apartado Dos muestras univariantes y una muestra bivariante). Es decir, en el primer caso se ha medido la variable A en unos sujetos de la población y la B en otros, mientras que en el segundo caso se coge la muestra y se miden las dos variables para cada sujeto. Volviendo a los gráficos, si formamos pares de valores con las dos variables (cuando tenga sentido, que no es siempre), pueden representarse como en el punto anterior, pero siempre teniendo en cuenta al interpretarlo qué información se ha unido, puesto que puede ser que haya diferencias entre los sujetos en que se ha medido A y los sujetos en que se ha medido B, y al final confundamos estas diferencias entre las muestras con diferencias entre las variables A y B. Este gráfico se parece más al descrito en (b) que al descrito en (a), y a veces es el único que se puede hacer.

Recta de regresión

El concepto de *regresión* fue propuesto por el biólogo Galton. Al estudiar la estatura de una muestra de padres e hijos, observó que los hijos de padres más altos que la media de la población tendían a tener una estatura mayor que la media pero menor que la de sus padres; del mismo modo, los hijos de padres más bajos que la media tenían hijos con una estatura menor que la de la media pero mayor que la de sus padres. En general se producía, por ambos lados, una «regresión a la media» en la estatura de una generación con respecto a la anterior.

Para intentar explicar el comportamiento medio de la relación entre dos variables se utiliza alguna función conocida que se le parezca. (Con «relación o comportamiento medio» se entiende «relación entre las medias».) La calidad de la modelización depende, en primer lugar, de la calidad de los datos.

Inspección visual

Como cualquier análisis de datos estadístico, es necesario empezar representando gráficamente (de cuantas más formas, mejor) los datos. La información de estos gráficos es fundamental. En este caso tiene una utilidad especial el diagrama de dispersión.

No se va a entrar en ciertos detalles ahora, sólo se mencionará que mediante la inspección visual se pueden identificar ciertos datos que tienen un efecto grande sobre el resultado final: *datos atípicos, puntos influyentes* o *puntos palanca*. A veces el efecto de estos datos se puede evaluar aproximadamente a partir del gráfico mirando si están alineados con el resto o no (en el caso de la recta de regresión), aunque también hay métodos analíticos para evaluar este efecto (haciendo los cálculos con y sin el punto que se quiere estudiar). Se incluye a continuación un gráfico, hecho por

una alumna a partir del fichero de datos «paises.txt», incluido con el libro *Introducción a la Estadística para las Ciencias Sociales*, de Daniel Peña y Juan-J. Romo.

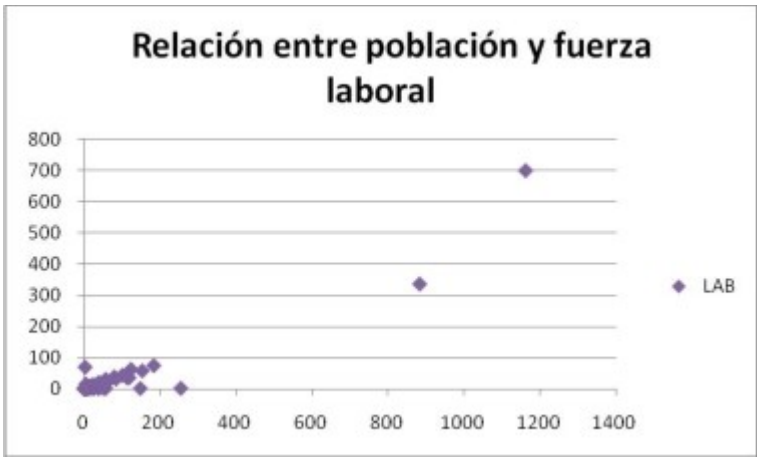
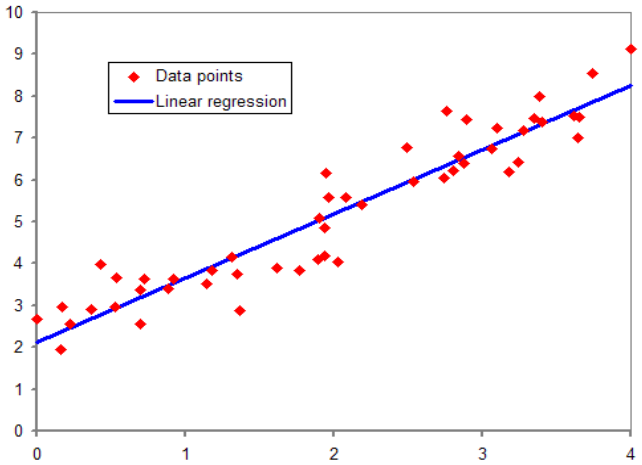


Figura: Muestra con dos puntos que pueden tener efecto en la recta de regresión

Una vez tratados los puntos especiales que se han mencionado, si los datos parecen estar situados a ambos lados de una recta imaginaria, tiene sentido elegir una recta (función lineal) para modelizar la dependencia de una variable con respecto a la otra. Esta modelización da lugar a la *recta de regresión*. Es decir, aunque matemáticamente siempre es posible calcular la recta de regresión, sólo debe hacerse cuando la inspección visual lo aconseje: véase lo que sucede con los datos en Correlación y relación no lineal entre variables. En el siguiente gráfico se muestra un ejemplo en que sí tiene sentido ajustar la recta de regresión.



http://es.wikipedia.org/wiki/Regresión_lineal

Obtención

De las posibles rectas cercanas entre sí que podrían utilizarse, se elige una mediante el criterio de hacer mínima la suma de las distancias, en vertical, de los puntos a la recta; para quitar el signo a estas distancias se las eleva al cuadrado, en vez de la opción menos manejable (para los cálculos) de tomar el valor absoluto. Por lo anterior, este criterio de elección de la recta de regresión se llama *método de mínimos cuadrados*. Si la muestra es $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, se quiere encontrar el modelo

$$Y = aX + b$$

(las letras a y b no tienen nada que ver con las variables A y B) tal que es mínima la cantidad

$$D=\sum_{i=1}^{n}(y_i-Y(x_i))^2=\sum_{i=1}^{n}(y_i-ax_i-b)^2.$$

Por ser una suma de cuadrados, nótese que D tomará el cero si, y sólo si, todas las cantidades de la suma son cero, es decir, cuando todos los puntos están perfectamente alineados (cosa que, obviamente, casi nunca pasará en la práctica). Por otro lado, dado que los valores de x_i y y_i son números concretos (los de la muestra), las cantidades que hay que elegir en el modelo son los parámetros a y b. Entonces, dejando «moverse» a los parámetros a y b, es decir, considerándolos como *variables* en vez de como *parámetros*, se elegirán sus valores adecuados concretos (con lo que volverán a ser parámetros); para ello la cantidad D anterior se interpreta como una función de dos variables, $D(a,b)$. Ahora, aplicando las técnicas del cálculo matemático para encontrar los extremos de una función de varias variables, se igualan a cero las derivadas parciales. Y como el extremo encontrado sólo puede ser un mínimo (porque D es mayor o igual a cero, y no está acotada por arriba: nos podemos imaginar una recta infinitamente lejos), entonces,

$$\begin{cases}\dfrac{\partial D(a,b)}{\partial a}=0\\[2ex]\dfrac{\partial D(a,b)}{\partial b}=0\end{cases}\quad\rightarrow\quad\begin{cases}a=\dfrac{s_{XY}}{s_X^2}\\[2ex]b=\bar{y}-a\cdot\bar{x}\end{cases}$$

donde $s_{XY}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$, $s_X^2=s_{XX}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$, $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ y $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

El proceso anterior nos ha llevado a que para calcular la recta de regresión es necesario utilizar las fórmulas obtenidas para a y b, y después sustituirlas en la expresión de la recta. (Esas fórmulas son las que tienen programadas las calculadores y los ordenadores para hacer estas rectas.)

Recta de *A* sobre *B* o recta de *B* sobre *A*

Al ajustar una recta de regresión, estamos modelizando una variable en función de la otra. Por tanto, es muy importante cuál es la variable independiente y cuál queremos estudiar en función de ella, esto es, la dependiente. El propio significado de las variables suele indicarlo. Por el proceso de construcción de la recta, que no trata igual a las dos variables (se miden las distancias en la dirección vertical), no se obtiene la misma recta al hacer la regresión de A en función de B que la regresión de B en función de A.

Representación

Una vez que hemos calculado los valores de a y b, hay varios métodos para representar la recta. Véase el apartado Representación a mano de una recta. En caso de que decidamos representarla dando dos valores a X, obtener los correspondientes valores de Y y representar estos dos puntos (dos puntos determinan una recta), los dos valores que elijamos para la variable independiente X deben estar en el rango de valores en que se mueve la muestra, porque la recta de regresión se ha construido con ella y puede no ser un modelo válido para valores lejanos. Los valores que eligiésemos, x_1 y x_2, pueden coincidir o no con dos valores que haya tomado la variable estadística A. Eso sí, si los tomamos para que coincidan es de esperar que los correspondientes y_1 y y_2 sean próximos a los valores que había tomado la variable estadística B en esos dos sujetos de la muestra, pero no tienen por qué coincidir (es muy difícil que lo hagan). La recta de regresión representa gráfica y analíticamente la relación que hay entre las medias de las variables estadísticas, pero en la realidad hay tantos factores influyendo en las medidas que los datos reales no estarán sobre la recta.

Por otro lado, si estuviésemos interesados en representar las dos rectas de regresión (de A sobre

B y de *B* sobre *A*) en un mismo gráfico, hay que tener en cuenta que una de ellas habría que «mirarla desde un lateral» del gráfico, puesto que su variable independiente estará en el eje vertical y no en el horizontal, como acostumbramos hacer en los ejes cartesianos.

Evaluación de la calidad

Una vez calculada la expresión de la recta de regresión, podemos estar interesados en evaluar su calidad a la hora de ajustarse a los datos, es decir, su calidad como modelo para explicar la relación media entre las variables. Una medida aproximada de esta calidad se obtiene a partir del gráfico, pero si se desea tener una medida analítica, ahora que conocemos los valores de *a* y *b* podemos volver a la expresión de *D* y obtener su valor numérico. Como se trata de una suma de *n* valores elevados al cuadrado, se definen los *residuos* como $r_i = y_i - Y(x_i) = y_i - ax_i - b$, esto es, las distancias en vertical de los datos de la muestra a los valores que predice el modelo; y se definen también la *varianza típica residual*

$$s_R^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - ax_i - b)^2 = \frac{1}{n} D$$

y la *desviación típica residual*

$$s_R = \sqrt{s_R^2} = \sqrt{\frac{1}{n} D}$$

En Ejercicio3 y Ejercicio 4 se calculan estas medidas.

Una observación que se puede hacer es que como *D* es mínima si, y sólo si, s_R^2 y s_R son mínimas, para una muestra dada la recta de regresión hace mínimas no sólo *D* sino también la varianza y la desviación típicas residuales.

Predicción

Si la calidad del modelo es buena, podemos utilizarlo para obtener valores estimados de la variable dependiente correspondientes a valores de la variable independiente que no estaban en su muestra: basta meter el valor de la *X* en la expresión de la recta. Como se ha indicado al hablar de la representación, el modelo es válido sólo en el rango de valores en que se mueve la muestra de la variable independiente: no tenemos garantías de que fuera de este rango la recta explica bien la relación entre las variables.

Relación con el coeficiente de correlación lineal

Es obvio que la recta de regresión, además de para explicar la relación media entre las variables y para predecir valores que no estén en la muestra, sirve para medir el grado de correlación lineal que hay entre las variables: cuanto mayor sea la correlación lineal, mejor ajustará la recta y menores serán los residuos. Se ha mencionado el caso límite en que $D=0$, lo que sucede si, y sólo si, los puntos están perfectamente alineados; nótese, a partir de sus expresiones, que en este caso también son cero la varianza y la desviación típicas residuales. Véase el apartado Relación entre la correlación y la recta de regresión.

Una expresión que relaciona la varianza típica residual, la varianza de la variable dependiente y el coeficiente de correlación lineal, *r*, es:

$$\frac{s_R^2}{s_Y^2} = 1 - r^2 .$$

(Esta letra r que se ha utilizado para el coeficiente de correlación no tiene nada que ver con la notación r_i de los residuos.)

Un poco más...

En esta sección se incluyen algunos comentarios de un nivel más alto, que pretenden por un lado facilitar la comprensión de lo visto y de los temas siguientes, y por otro satisfacer la curiosidad de quien quiera ir un poco más allá de lo estrictamente necesario según el temario.

Ecuaciones de la recta

Quien necesite –o quiera– repasar las ecuaciones de la recta, puede consultar el apartado correspondiente y los enlaces que se dan en el documento

«Repaso no del todo elemental de Matemáticas elementales»

http://www.Casado-D.org/edu/RepasoMatematicas.pdf

Representación a mano de una recta

En el fondo, una función matemática de la forma $y = f(x)$ es una regla que «informa a la vez de infinitos pares de valores (x,y)».

Para representar una función hay varias opciones, una sencilla es construir una tabla con una cantidad suficiente de estos pares de puntos, representarlos en un gráfico y unirlos suavemente (si la función debe ser suave).

En el caso de la recta, una primera forma muy poco eficaz para representarla sería ajustar la recta de regresión a los puntos de esta tabla mencionada en el párrafo anterior. Aquí la recta debe ajustar perfectamente y pasar por todos los puntos (su desviación residual debe ser cero). Por supuesto, éste es un camino que no utiliza nadie, es como «matar moscas a cañonazos»; pero pensar en él ayuda a comprender mejor los conceptos.

Volviendo a la recta, por tratarse de una ecuación lineal, dos puntos la determinan (como curiosidad: existen en Matemáticas otros tipos de *geometrías no euclídeas* en que por dos puntos pueden pasar más de una «recta» distintas). Por tanto, lo más cómodo es dar dos valores a la X y obtener sus valores de Y. Ahora ya se pueden representar los puntos (x_1,y_1) y (x_2,y_2), para unirlos finalmente con una regla. En teoría, si la ecuación de la recta está ya calculada, daría igual dónde coger los dos puntos, porque determinan igual el gráfico de la recta. Sin embargo, dado que en los dibujos a mano a veces marcamos aproximadamente los valores, el gráfico queda mejor tomando puntos parecidos a los de la muestra. Si se hace con ordenador, si no se tomasen puntos parecidos a

los de la muestra sucedería que los puntos de la muestra aparecerían ocupando una parte muy pequeña del gráfico, lo que haría al gráfico perder mucha utilidad y precisión. Es decir, en este caso se toman valores parecidos a los de la muestra por conveniencia, mientras que en Recta de regresión se tomaban estos valores cercanos a los de la muestra no sólo por este motivo de conveniencia sino por necesidad (validez del modelo).

Otra forma de representar la recta es a partir de las distintas formas que tienen sus ecuaciones. Dependiendo de cómo esté dada la información, se utilizará una u otra forma. De entre estas ecuaciones, la más sencilla es la *ecuación punto-pendiente*. Quien tenga interés en estas ecuaciones de la recta puede consultar el apartado correspondiente del documento

«Repaso no del todo elemental de Matemáticas elementales»

http://www.Casado-D.org/edu/RepasoMatematicas.pdf

Plantear una ecuación del enunciado

Hay casos en que no nos dan explícitamente la relación entre las variables, tenemos que plantearla a partir del enunciado del ejercicio.

Podemos llamar como queramos a lo que queramos, siempre que después seamos consecuentes con el concepto que hay detrás del nombre. Suele llamarse X a la variable independiente e Y a la dependiente, la que depende de la X. Seguir este convenio para nombrar a las variables ayuda mucho a pensar, por eso la notación es importante en Matemáticas, sobre todo al principio. Después uno puede pensar que Y es la independiente y X la dependiente, no pasa nada, pero no ayuda a dedicar esfuerzos mentales a lo realmente importante.

Por ejemplo, si tenemos las variables, «ruido» y «número de lectores», ¿cuál puede depender de la otra?... Pues parece que el ruido depende del número de lectores, más que al revés.

A la hora de hacer un planteamiento hay que:

- Definir bien y claramente las variables: X será el número de lectores e Y será el nivel de ruido.
- Establecer la ecuación según el enunciado, por ejemplo: $5Y-10X = 20$
- Despejar la variable dependiente: $Y = (20+10X)/5 = 20/5 + 10X/5 = 4 + 2X$

Vemos que la relación entre estas dos variables es una recta con pendiente 2. Además, cuando $X = 0$ se tiene que $Y = 4$, así que hay ruido en la biblioteca aunque no haya lectores (luces, calefacción, ratones de biblioteca, espíritus y duendes...).

Resolución gráfica de algunos problemas

Algunos ejercicios pueden ser resueltos tanto de forma analítica como de forma gráfica. Por ejemplo, pongamos el caso en que después de plantear las ecuaciones para los precios de dos bibliotecas en función del número de libros, se obtienen:

Biblioteca A: $Y_A = 3X + 12$ Biblioteca B: $Y_B = 9X$

Estas ecuaciones ya nos están dando mucha información analíticamente:

- Cuando no sacamos ningún libro, es decir, X=0, en la biblioteca A tenemos que pagar 12€, mientras que en la B no pagamos nada.
- Si nos preguntamos para cuántos libros tenemos que pagar lo mismo en ambas bibliotecas, basta resolver el sistema formado por las dos ecuaciones a la vez, es decir:

$$Y_A = Y_B \rightarrow 9X = 3X + 12 \rightarrow 9X - 3X = 12 \rightarrow 6X = 12 \rightarrow X = 12/6 = 2$$

- Podemos definir la función «diferencia», D, que nos diga la diferencia entre ambos precios:

$$D(X) = Y_A(X) - Y_B(X) = 3X + 12 - 9X = -6X + 12$$

  Esta función informa, para cada valor de X, de cuánto más hay que pagar en A que en B. Cuando esta función toma valores negativos significa que en B se paga más que en A.

Hemos llamado X a la variable «número de libros» e Y a la variable «euros». Por su significado, vemos que la variable Y depende de la variable X, es decir, la primera es la dependiente y la segunda la independiente. Damos estos nombres porque nos va a facilitar el pensar, pero debemos tener en cuenta que en este caso las dos variables son discretas y van a tomar sólo valores naturales. Podríamos haber elegido la notación

Biblioteca A: $e_A(n) = 3n + 12$ Biblioteca B: $e_B(n) = 9n$

y tendríamos los euros escritos en función del número de libros. Parece que la n nos ayuda a pensar en valores naturales. En cualquier caso, es una cuestión de nombres.

La solución más fácil al problema es hacer una tabla como la siguiente:

Libros	Biblioteca *A*	Biblioteca *B*	Coste
1	15	9	-6
2	18	18	0
3	21	27	6
4	24	36	12
5	27	45	18

Por otro lado, dado que las ecuaciones de las bibliotecas se pueden representar como puntos de rectas, podríamos hacer el siguiente gráfico:

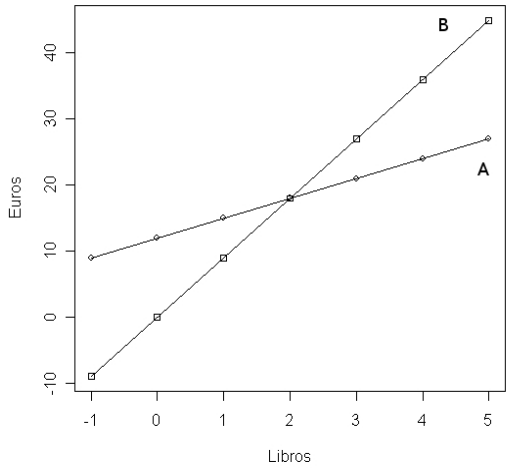


Figura: Rectas que representan los precios de las bibliotecas

De él se ve inmediatamente que:

- Para menos de dos libros, el precio en *B* es menor que en *A*.
- Para dos libros, ambos precios son iguales
- Para muchos libros, los precios de la biblioteca de menor pendiente son menores.

Otros tipos de regresión

En el apartado Recta de regresión se ha hablado de la modelización entre dos variables que parecen tener una dependencia, en media, lineal. Se trataba, por tanto, de la *regresión lineal simple*. Sin embargo, la idea principal, esto es, minimizar la suma de las distancias al cuadrado de la muestra a un modelo (método de mínimos cuadrados), es perfectamente aplicable para intentar explicar otros tipos de relaciones entre variables.

Regresión (no lineal) cuadrática

A la vista de los datos del gráfico de Correlación y relación no lineal entre variables, se podría utilizar el criterio de los mínimos cuadrados para encontrar la expresión cuadrática que mejor se adapte a los datos. En este caso, si la muestra es $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, se quiere hacer una *regresión no lineal cuadrática*, es decir, encontrar el modelo $Y = aX^2 + bX + c$ tal que es mínima la cantidad

$$D=\sum_{i=1}^{n}(y_i-Y(x_i))^2=\sum_{i=1}^{n}(y_i-ax_i^2-bx_i-c)^2$$

De nuevo D se puede interpretar como una función de varias variables, $D(a,b,c)$, lo que llevaría al problema matemático de calcular las derivadas y resolver el siguiente sistema de tres ecuaciones con tres incógnitas:

$$\begin{cases} \dfrac{\partial D(a,b,c)}{\partial a}=0 \\ \dfrac{\partial D(a,b,c)}{\partial b}=0 \\ \dfrac{\partial D(a,b,c)}{\partial c}=0 \end{cases} \rightarrow \begin{cases} a=\cdots \\ b=\cdots \\ c=\cdots \end{cases}$$

Regresión lineal múltiple

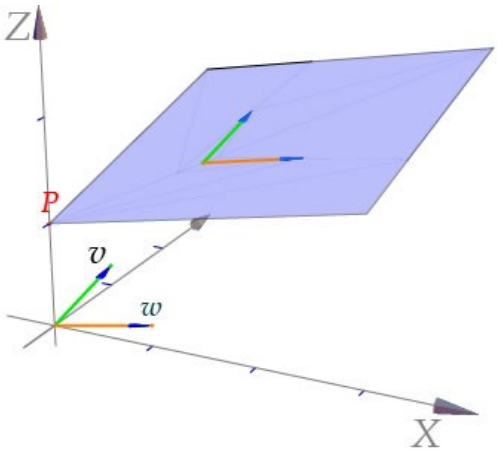
Podemos tener interés en modelizar cómo depende una variable Z de varias variables, pongamos X e Y. En este caso, si la muestra es $(x_1,y_1,z_1),(x_2,y_2,z_2),\ldots,(x_n,y_n,z_n)$, para hacer una *regresión lineal múltiple*, se quiere encontrar el modelo $Z = aX + bY + c$ tal que es mínima la cantidad

$$D=\sum_{i=1}^{n}(z_i-Z(x_i,y_i))^2=\sum_{i=1}^{n}(z_i-ax_i-by_i-c)^2$$

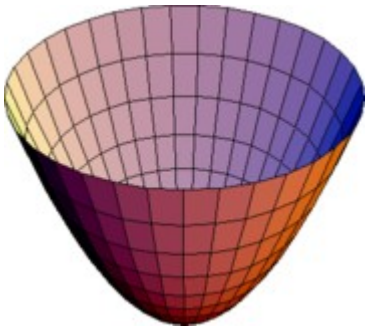
Las distancias se miden ahora se miden «en altura» desde los puntos al plano. De nuevo D se puede interpretar como una función de varias variables, $D(a,b,c)$, lo que llevaría al problema matemático y la solución:

$$\begin{cases} \dfrac{\partial D(a,b,c)}{\partial a}=0 \\ \dfrac{\partial D(a,b,c)}{\partial b}=0 \\ \dfrac{\partial D(a,b,c)}{\partial c}=0 \end{cases} \rightarrow \begin{cases} a=\cdots \\ b=\cdots \\ c=\cdots \end{cases}$$

Es fácil imaginarse este tipo de regresión en un gráfico. Generalizando el dibujo que se ha incluido para la regresión lineal simple, ahora hay una serie de puntos «flotando» en el espacio tridimensional y, como parece que se sitúan a ambos lados de un plano, tiene sentido buscar el plano que mejor ajusta a los datos. De la misma manera que la recta de regresión informaba de lo linealmente que Y dependía de X (a través de la *correlación lineal*), ahora el plano de regresión informa de lo linealmente que Z depende de X e Y (a través del concepto de *correlación múltiple*). También se podría generalizar la regresión cuadrática anterior a una *regresión cuadrática múltiple* o *paraboloidal* (estos nombres son míos, no sé si en la literatura se nominan así estos conceptos), de la misma forma que se ha generalizado la lineal simple a la múltiple.



http://www.cidse.itcr.ac.cr/cursos-linea/SUPERIOR/t2-Funciones-de-variasvariables/3-algebra-vectorial/node5.html



http://es.wikipedia.org/wiki/Paraboloide

Para este problema de modelizar datos relacionados no linealmente, otra aproximación diferente consistiría en aplicar alguna transformación a los datos para que la dependencia entre los datos transformados pase a ser lineal y se pueda ajustar el modelo lineal; por último, después de ajustado el modelo, no habría que olvidarse de deshacer la transformación que se hubiese aplicado.

Cálculo de medidas numéricas

Correlación y relación no lineal entre variables

Diagrama de dispersión

Como se ve claramente en el siguiente gráfico de dispersión de dos variables, la relación media que hay entre estas variables es cuadrática (por la forma de parábola), no lineal.

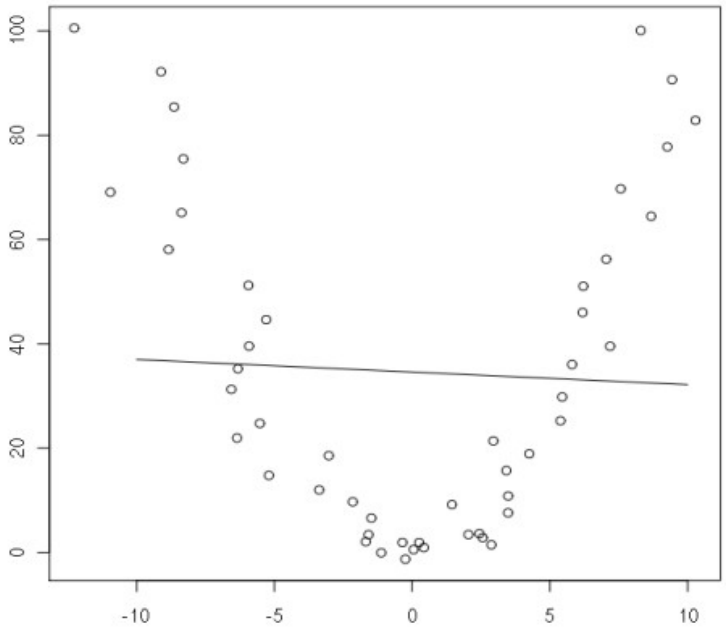


Figura: Dos muestras con relación media cuadrática y su recta de regresión

Matemáticamente, una parábola tiene la forma $Y = aX^2 + bX + c$. Vemos que tiene un término más que una recta: aX^2 es el *término cuadrático*, bX es el *término lineal* y c es el *término independiente*.

Correlación lineal

Si calculásemos el valor numérico de la correlación lineal entre las dos variables, veríamos que es bastante cercana a cero. Sin embargo, a simple vista se ve claramente que las variables están muy relacionadas. ¿Qué está pasando aquí? Lo que sucede es que la covarianza y la correlación sólo miden la relación lineal entre dos variables, no «ven» otro tipo de relación, no «ven» que las variables están relacionadas cuadráticamente.

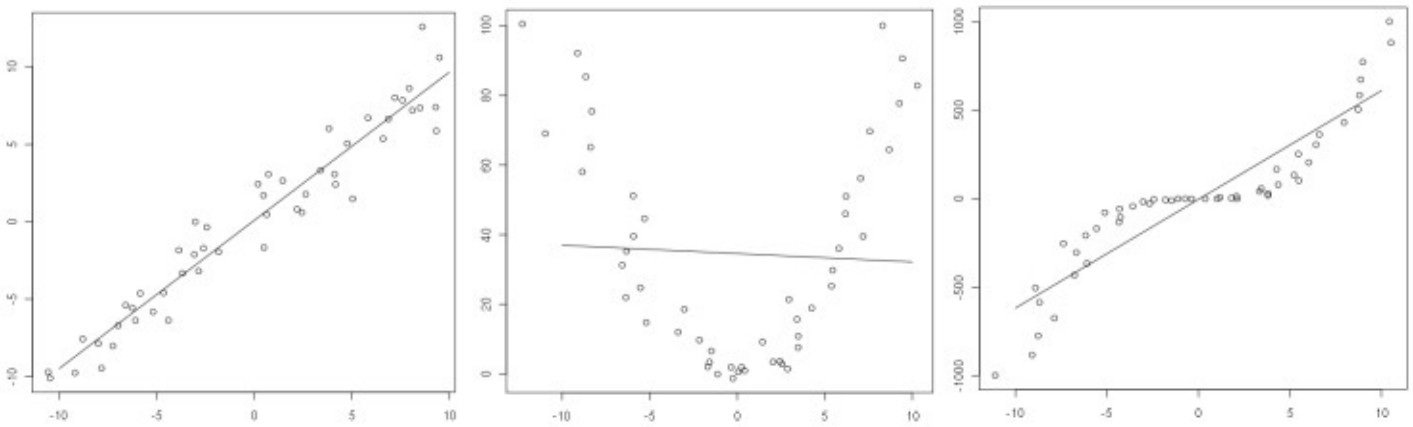
Recta de regresión

La recta de regresión siempre se puede hallar, pero no siempre se debe. Es decir, se ve en el dibujo que la recta no representa bien la relación media entre las variables. En casos así lo correcto sería coger la fórmula de una parábola y, siguiendo la misma idea que en la recta de regresión de minimizar la suma de las distancias verticales de los puntos a la curva, hallar los coeficientes que determinarían cuál es esa parábola concreta que mejor representa la relación entre las variables.

En este conjunto de datos, el que la recta obtenida aquí tenga una pendiente casi nula significaría que $Y = aX + b \approx b$, es decir, casi desaparece la X en la expresión de la Y. Si vemos la recta como la relación entre las medias de las variables, esto estaría diciendo que la media de la variable Y tiende a ser siempre b, independientemente de cuánto valga la media de X. Más profundamente, el que la media de Y sea b lo que significa es que sus valores tienden a distribuirse con su media en b; esto es, con igual «masa» a ambos lados de ese valor: puede haber los mismos datos, en tamaño y distancia, a ambos lados, o menos datos pero más distantes a un lado.

La pendiente informa de la correlación entre las dos variables. Éste es un detalle importante, se explica con más detalle en el apartado Relación entre la correlación y la recta de regresión.

En la figura siguiente se muestran también los casos de relaciones lineal y cúbica. Sólo en el caso lineal la recta de regresión es un modelo correcto de la relación media entre las variables. En el caso de la relación cuadrática hemos mencionado que la recta es un modelo muy malo. Finalmente, en el caso de la relación cúbica, sin ser un modelo correcto, el error es menor que en el caso cuadrático (por la forma de la propia función cúbica).



$Y = X$ $\qquad$ $Y = X^2$ $\qquad$ $Y = X^3$

Figura: Tres muestras bivariantes, sus rectas de regresión y, abajo, el modelo correcto de relación media entre las variables

Relación entre la correlación lineal y la recta de regresión

Conviene insistir en la relación que hay entre el coeficiente de correlación lineal y la recta de regresión.

Coeficiente de correlación lineal

Este coeficiente se define a partir de la covarianza, $cov(X,Y)=s_{XY}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})$, como:

$$r=\frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}}=\frac{s_{XY}}{s_X s_Y}$$

donde recordemos que la varianza para cada variable se obtenía como la covarianza de ella consigo misma $var(X)=s_X^2=s_{XX}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2$.

Recta de regresión

Al hacer los cálculos de la recta de regresión $Y=aX+b$ se llega a que:

$$a=\frac{s_{XY}}{s_X^2}.$$

Relación

Por tanto, podemos hacer

$$a=\frac{s_{XY}}{s_X^2}=\frac{s_{XY}}{s_X^2}\cdot\frac{s_Y}{s_Y}=\frac{s_{XY}}{s_X s_Y}\cdot\frac{s_Y}{s_X}=r\frac{s_Y}{s_X}.$$

Y ésta es la relación entre la pendiente de la recta de regresión y el coeficiente de correlación. Vemos que una tiende a ser pequeña si, y sólo si, la otra lo es. Esto es lo que pasaba en el caso visto en el apartado Correlación y relación no lineal entre variables: tanto la pendiente como la correlación eran próximas a cero. Sin embargo, el cociente s_Y/s_X también tiene su importancia, porque podría haber casos en que fuese muy grande o muy pequeño, dependiendo de las desviaciones típicas de ambas muestras.

Un poco más...

En esta sección se incluyen algunos comentarios de un nivel más alto, que pretenden por un lado facilitar la comprensión de lo visto y de los temas siguientes, y por otro satisfacer la curiosidad de quien quiera ir un poco más allá de lo estrictamente necesario según el temario.

Inestabilidad numérica

El Cálculo Numérico, como rama de las Matemáticas, trabaja con métodos y algoritmos de cálculo cuyo resultado final son números, no objetos matemáticos abstractos como variables, funciones, funcionales o conjuntos.

En las operaciones con números, sobre todo con las facilidades que aportan los ordenadores hoy en día, son importantes conceptos como: mantisa, cifras decimales, truncamiento, redondeo, errores, etcétera.

Al resolver un ejercicio en que había que ajustar una recta de regresión a la siguiente muestra de datos bivariantes

A	2215	2518	1781	1823	2117	2703	1423	1532	1879	1713	2122	2346
B	58,3	62,9	46,3	48,2	58,2	65,8	36,7	32,3	52,7	39,3	58,7	39,3

dos alumnas obtenían, respectivamente

Alumna 1	Alumna 2
$Y - 49{,}89 = 0{,}02(X - 2014{,}33)$	$Y - 49{,}89 = 0{,}022(X - 2014{,}3)$
$Y = 49{,}89 + 0{,}02(X - 2014{,}33)$	$Y = 49{,}89 + 0{,}022(X - 2014{,}3)$
$Y = 49{,}89 - \mathbf{40{,}29} + 0{,}02X$	$Y = 49{,}89 - \mathbf{44{,}3146} + 0{,}022X$
$Y = 0{,}02X + \mathbf{9{,}6}$	$Y = 0{,}022X + \mathbf{5{,}57}$

Al buscar la causa de esta diferencia y hacer los siguientes cálculos con un lenguaje de programación (**R**, en este caso)

```
> 0.02*2014.33
[1] 40.2866

> 0.022*2014.3
[1] 44.3146
```

salió a la luz que la diferencia, al haber tomado distinto tipo de cifras decimales, se debía al problema de inestabilidad numérica que aparecía al trabajar con un número tan pequeño como es 0,02... En general, al hacer operaciones con números muy cercanos a cero hay que tener mucho cuidado. En este caso, esa cantidad proviene, según la fórmula de la recta de regresión ($Y = aX + b$), de

$$a = \frac{s_{AB}}{s_A^2} = 0{,}02...$$

Es decir, que al tomar a la variable A como variable independiente, por ser sus valores de un orden mucho mayor que los de la variable B, el coeficiente de la recta de regresión es muy pequeño y crea un problema de inestabilidad numérica. El problema se soluciona midiendo A en otra unidad.

La diferencia entre los términos independientes de los modelos de ambas alumnas es aproximadamente de 9,6-5,6 = 4 unidades, cantidad que no es pequeña comparada con los valores de la muestra B (su media es 49,89), que es la variable dependiente. Nótese que los términos independientes son valores que siempre están afectando a la variable dependiente en la expresión de la recta. Podemos concluir que el problema de inestabilidad no es despreciable.

Simulaciones

Cuando queremos comprobar cómo se comporta un determinado método que trabaja con datos, se conoce como *simulación* al hecho de:

1. Generar unos datos a partir de un modelo concreto conocido.
2. Hacer como que no conocemos el modelo, es decir, tratar los datos como si hubiesen sido recogidos de una muestra realmente. Introducir estos datos en el método que queremos probar, sin indicarle de qué modelo provienen los datos.
3. Como en realidad sí conocemos el modelo que había generado los datos, podemos evaluar cuán bien se ha comportado el método con ese modelo de datos.

La idea es sencilla. (La Ciencia y la Vida están llenas de ideas sencillas... que nunca nadie tiene tiempo de explicarnos.) Las simulaciones se utilizan mucho en Economía, Ingeniería o Estadística, por ejemplo. En concreto, en Estadística podemos querer saber cómo se comporta un método a la hora de estimar medidas de las que hemos visto en la teoría. En vez de intentar conseguir cientos de muestras de datos reales para probar el método en un amplio número de casos/modelos distintos, lo que sería difícil y costoso, con las simulaciones podemos ver fácilmente qué haría el método si realmente le «trajésemos de la calle» muestras de esos modelos concretos (no entramos en el hecho de que los datos reales son muchas veces más complicados que los datos simulados).

Esto es lo que se ha hecho para generar los datos que se han utilizado en el apartado Correlación y relación no lineal entre variables:

- Para la variable *A*: Se han generado posiciones equidistantes entre -10 y +10.
- Para la variable *B*:
  - Se han cogido esos datos y se les ha aplicado la fórmula $b_i = a_i^2$ a cada uno, de manera que se obtienen puntos sobre una parábola perfecta con su mínimo en el $x = 0$.
  - Y ahora, para «esconder» un poco esa parábola, se ha añadido a cada dato un ruido/error aleatorio, generado con una variable aleatoria hacia arriba o hacia abajo indistintamente.
- Finalmente, para «esconder» también un poco la variable *A*, le he añadido a cada dato un ruido o error aleatorio, generado con una variable aleatoria hacia uno u otro lado indistintamente.

El resultado es una nube de puntos que tienden a situarse alrededor de la parábola. Todos estos cálculos se pueden hacer con cualquier lenguaje de programación o con el generador de números aleatorios de *Excel*.

Comprender las simulaciones ayuda a comprender gran parte de la filosofía que hay detrás de la Estadística. La idea de los temas de Probabilidad es ver los modelos teóricos que se utilizan para intentar modelizar las variables estadísticas. Es decir, la Estadística frecuentemente supone que los datos de una muestra de datos reales se generan con alguna regla determinista en principio no conocida a la que, por diversos motivos y factores, se le añade un error aleatorio. Nosotros lo que observamos son estos resultados finales (factor determinista más error aleatorio) e intentamos averiguar el factor determinista (que nos dirá cómo dependen las acciones de una empresa, cómo depende la resistencia de un material del agua utilizada en su fabricación, etcétera) y las características de los errores aleatorios.

Apuntes sobre dependencia y correlación lineales

En general, hay que tener claro cuándo los términos «dependiente» y «independiente» se están utilizando en sentido estadístico-probabilístico o en sentido lingüístico genérico, porque en este último siempre que una variable se pone en función de otra, sea linealmente o no, ambas se están relacionando y se suele decir también que son «dependientes». Incluso dentro de otras ramas de la Ciencia más alejadas se pueden encontrar a la vez ejemplos de los dos usos anteriores.

Se hacen ahora dos pequeños apuntes sobre los términos «dependiente» e «independiente», cuando se utilizan con sentido estadístico-probabilístico. Como ya se ha dicho, el coeficiente de correlación (lineal) sólo mide el grado de relación lineal entre dos variables, lo que indica cómo deben entenderse aquí estos términos.

Ausencia total de correlación

Con las definiciones de independencia y correlación para variables aleatorias, se cumple que:

$$X \text{ e } Y \text{ son independientes} \quad \Rightarrow \quad \rho(X,Y)=0$$

Sin embargo, la implicación inversa es sólo cierta si las variables aleatorias siguen una distribución normal:

$$\rho(X,Y)=0 \quad \text{y } X \text{ e } Y \text{ son normales} \quad \Rightarrow \quad X \text{ e } Y \text{ son independientes}$$

Presencia total de correlación

Este otro apunte, un poco más difícil de interpretar, se relaciona con que la correlación alcanza su valor máximo (salvo el signo). Al hablar de *rectas* en Análisis matemático, los términos «dependiente» e «independiente» significan esencialmente lo mismo que al hablar de *correlación lineal* en Estadística y Probabilidad (y aquí hay dos definiciones, una para variables estadísticas y otra para variables aleatorias), aunque no lo parezca. Se puede decir que cuando una ecuación lineal relaciona dos variables su correlación es uno. En el libro *Probability and Random Processes*, de G. Grimmett y D. Stirzaker, se puede encontrar que, para variables aleatorias (lema 8 del apartado 3.6 para las discretas y este mismo lema junto con el teorema 12 del apartado 4.5 para las continuas):

$$|\rho(X,Y)|=1 \quad \Leftrightarrow \quad P(a_1X+a_2Y=a_3)=1, \quad a_1,a_2,a_3\in\mathbb{R}$$

Es decir, la correlación es uno en valor absoluto (luego el signo + o – está relacionado con que la relación entre las variables sea directa o inversa) si, y sólo si, la relación entre las variables es lineal «en el mundo de la probabilidad» (técnicamente se habla de «con probabilidad uno»). Otro aspecto de matemáticas avanzadas, del que no se va a hablar ahora, es cuándo una expresión literal (de letras) que relaciona dos variables, por ejemplo

$$Y=aX+b$$

debe entenderse que expresa una relación totalmente determinista, en casi todo punto, probabilísticamente, etcétera.

Algunos ejercicios resueltos

Ejercicio 3

Como parte de la evaluación de un sistema de computadoras, un gerente de sistemas quiere predecir el tiempo de respuesta de las terminales de computadora. El tiempo de respuesta de una terminal se define como el tiempo (T), en segundos, que la computadora tarda en responder a un comando enviado desde una terminal. Aunque son muchas las variables que influyen en el tiempo de respuesta de las terminales, el gerente de sistemas va a modelizar dicho tiempo como una función del número de usuarios simultáneos (N) en el momento en que envía el comando. El gerente ha recabado los datos de estas dos variables en cinco instantes de tiempo, de forma que:

$$\sum_{i=1}^{5} n_i = 15, \quad \sum_{i=1}^{5} t_i = 4{,}60, \quad \sum_{i=1}^{5} n_i \cdot t_i = 16{,}97, \quad \sum_{i=1}^{5} n_i^2 = 55 \ \ y \ \ \sum_{i=1}^{5} t_i^2 = 5{,}28$$

a) *¿Existe una relación lineal entre las variables? Usar algún coeficiente que indique cómo es esta relación.*

b) *Calcular la recta de regresión, que relaciona el tiempo de respuesta de una terminal con el número de usuarios simultáneos en ese momento.*

c) *Interpretar los parámetros del ajuste de regresión.*

d) *Razonar si es posible que la varianza residual tome el valor de 0,30.*

La primera observación que hacemos es que no disponemos de los valores originales de las muestras de las dos variables estadísticas; a cambio nos dan como información algunas cantidades calculadas a partir de estas muestras. Podríamos preguntarnos si se pretende que recuperemos las muestras a partir de estas informaciones, pero pronto nos damos cuenta de que la información del enunciado constituye un conjunto de cinco ecuaciones (algunas de ellas no lineales, puesto que tienen exponentes mayores a uno), mientras que el número de incógnitas sería diez (los cinco valores de cada muestra). Por tanto, parece que no es éste el camino que tenemos que tomar para resolver el ejercicio. Parece que tendremos que calcular las cantidades que nos pidan expresándolas en función de la información que nos dan en el enunciado, que tiene que ser *suficiente*.

a) Nosotros medimos la correlación lineal principalmente con el coeficiente de correlación lineal, que se define como $r := \frac{s_{XY}}{s_X \cdot s_Y}$. Con los nombres de nuestras variables quedaría:

$$r := \frac{s_{NT}}{s_N \cdot s_T}$$

Necesitamos calcular la covarianza y las desviaciones típicas que aparecen en la fórmula.

- Para calcular la covarianza entre dos variables estadísticas tenemos, en general, dos

expresiones: $Cov(X,Y)=\frac{1}{m}\sum(x_i-\bar{x})(y_i-\bar{y})$, que corresponde a la definición, y $Cov(X,Y)=\frac{1}{m}\sum x_i y_i-\bar{x}\,\bar{y}$, que se deduce de ella como una propiedad. En cada ejercicio utilizaremos la que podamos o nos resulte más apropiada. Por la información del enunciado, ahora no nos queda más remedio que utilizar la segunda fórmula, puesto que no conocemos los valores concretos de las muestras de las variables:

$$Cov(N,T)=\frac{1}{m}\sum n_i t_i-\bar{n}\,\bar{t}=\frac{1}{5}16{,}97-\frac{1}{5}15\cdot\frac{1}{5}4{,}6=3{,}394-2{,}76=0{,}634$$

- Para calcular la varianza también existen, en general, dos fórmulas: $Var(X)=\frac{1}{m}\sum(x_i-\bar{x})^2$, que es la definición, y la siguiente, que se deduce como propiedad, $Var(X)=\frac{1}{m}\sum x_i^2-\left(\frac{1}{m}\sum x_i\right)^2=\frac{1}{m}\sum x_i^2-(\bar{x})^2$. De nuevo, en nuestro caso, sólo podemos utilizar esta segunda:

$$s_N^2=\frac{1}{5}\sum n_i^2-\left(\frac{1}{5}\sum n_i\right)^2=\frac{1}{5}55-\left(\frac{1}{5}15\right)^2=11-9=2 \quad\rightarrow\quad s_N=\sqrt{s_N^2}=\sqrt{2}$$

$$s_T^2=\frac{1}{5}\sum t_i^2-\left(\frac{1}{5}\sum t_i\right)^2=\frac{1}{5}5{,}28-\left(\frac{1}{5}4{,}6\right)^2=1{,}06-0{,}85=0{,}21 \quad\rightarrow\quad s_T=\sqrt{s_T^2}=\sqrt{0{,}21}$$

Por tanto, concluimos que $r:=\frac{s_{NT}}{s_N\cdot s_T}=\frac{0{,}634}{\sqrt{0{,}42}}=0{,}98$. Es decir, existe relación lineal entre las variables N y T, esta relación es positiva (cuanto mayor es una variable mayor es la otra) y muy fuerte (dado lo cerca que r está de 1, su valor máximo).

<u>Observación</u>: Como su nombre indica, el coeficiente de correlación lineal sólo es capaz de encontrar si existe o no relación lineal entre dos variables, pero no es capaz de descubrir relaciones no lineales: por ejemplo, si dos variables están relacionadas cuadráticamente por $Y = aX^2 + bX + c$. Sobre esto se ha hablado en Correlación y relación no lineal entre variables.

b) En este caso, está claro que en la regresión N debe ser la variable independiente y T la dependiente. En general, es muy importante tener claro cuál es la variable dependiente y cuál la dependiente. Para calcular la recta vamos a calcular los coeficientes a partir de las expresiones teóricas:

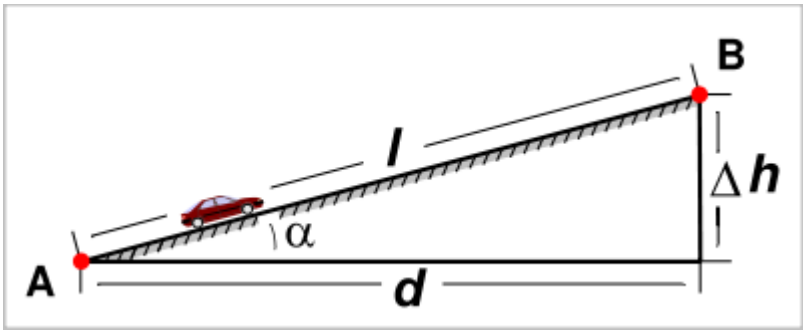
- $a=\frac{s_{N,T}}{s_N^2}=\frac{0{,}64}{4}=0{,}32$ (Nótese que la varianza del denominador es la de la variable independiente.)

- $b=\bar{t}-a\,\bar{n}=\frac{1}{5}4{,}6-0{,}32\frac{1}{5}15=-0{,}04$

La recta de regresión es, entonces: $T = aN + b = 0{,}32N - 0{,}04$. Es fácil hacer una

representación gráfica en un plano de coordenadas.

c) Como en toda recta, en la de regresión lineal el coeficiente de la variable indica la pendiente. Esta pendiente puede verse como:



$$a = tg\,\alpha = \frac{\Delta h}{d} = \frac{a}{1}$$

Es decir, podemos imaginar un triángulo cuyo cateto contiguo mida 1 y cuyo cateto opuesto mida *a*, de donde interpretamos que *a* es la cantidad que aumenta la variable dependiente cuando la variable independiente aumenta en una unidad. Como en nuestro caso la pendiente de la recta es 0,32, cuando aumentamos un usuario simultáneo en el momento de enviar un comando, el tiempo de respuesta de una terminal aumenta, en media, 0,32 segundos.

El término independiente de la ecuación de una recta es el punto en el que la recta corta al eje vertical, o, lo que es lo mismo, el valor que toma la variable dependiente cuando la independiente toma el valor 0. En este caso, el punto donde la recta corta al eje de ordenadas es *N*=0, *T*=-0,04. Respecto a este valor, y aunque es prácticamente cero, siendo exhaustivos puede parecer que refleja un pequeño error del ejercicio, dado que la media de una variable positiva (tiempo) no puede ser negativa. No obstante, dado que no conocemos los valores de la muestra, no sabemos si la recta es un modelo que pueda aplicarse a valores cercanos a N=0 (una recta es válida como modelo para el rango de valores de la variable independiente que abarque su muestra).

d) La varianza típica residual es una medida que evalúa el ajuste de la recta a los valores de la muestra; es decir, evalúa la calidad del modelo. Se define como $s_R^2 = \frac{1}{m}\sum (t_i - \hat{t}_i)^2$, donde t_i es el *i*-ésimo valor de la muestra de *T* y $\hat{t}_i$ es el valor estimado para n_i utilizando la recta de regresión, esto es, $\hat{t}_i = 0{,}32\,n_i - 0{,}04$. No podemos utilizar la expresión de la definición anterior, por lo que tenemos que pensar en otra fórmula que involucre la varianza típica residual; esta fórmula es:

$$\frac{s_R^2}{s_T^2} = 1 - r^2,$$

de donde, $s_R^2 = (1 - r^2)\cdot s_T^2 = (1 - 0{,}96)\cdot 0{,}21 = 0{,}01$. Vemos que la varianza residual no toma el valor por el que se preguntaba en el enunciado.

Ejercicio 4

Tenemos cinco alumnos de un programa de apoyo escolar en la asignatura de Matemáticas. Se pretende establecer una relación entre los resultados académicos (obtenidos por el orientador escolar a través de una prueba psicopedagógica) que obtuvieron antes de entrar al programa (A) y los obtenidos ahora (B). Los resultados son los siguientes:

A	a_1	-4	0	3	5
B	3	1	2	6	5

En el caso de que a_1 *tome el valor de -1:*

a) *¿Cuál de las dos medidas está más dispersa?*

b) *Determina y dibuja la recta de regresión de B sobre A, así como el coeficiente de correlación. ¿Cómo se interpreta el resultado del coeficiente de correlación?*

c) *Determina los resultados esperados para un alumno que se incorporó más tarde al programa y que obtuvo como evaluación inicial 2.*

d) *¿Qué relación existe entre el coeficiente de regresión de B sobre A y el coeficiente de correlación?*

e) *Calcula los residuos obtenidos y la desviación típica residual.*

f) *El psicopedagogo ha traspapelado el valor de* a_1*, pero recuerda que la desviación típica de la variable A coincidía con su coeficiente de variación. ¿Cuál es el valor de* a_1*?*

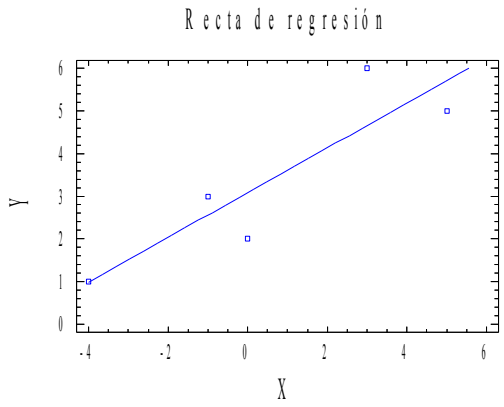
a) La medida de dispersión que se utiliza para comparar dos poblaciones distintas es el coeficiente de variación, puesto que tiene en cuenta tanto la localización (a través de la media) como la variabilidad (a través de la desviación típica) de los valores dentro de cada población. Se define como:

$$CV_A = \frac{s_A}{|\bar{a}|}.$$

En este caso, $CV_A = 5{,}228$ y $CV_B = 0{,}545$, por lo que la variable *A* es más dispersa (sus valores se distribuyen con mayor dispersión).

b) Haciendo cálculos se obtiene que: $\bar{a} = 0{,}6$, $\bar{b} = 3{,}4$, $s_B^2 = 3{,}44$, $s_{AB} = 5{,}16$ y $r = 0{,}8872$. Queremos hallar $Y = mX + n$. Calculamos los coeficientes de la recta con $m = \frac{s_{AB}}{s_A^2} = \frac{5{,}16}{9{,}84} = 0{,}52$ y $n = \bar{b} - m\,\bar{a} = 3{,}4 - 0{,}52 \cdot 0{,}6 = 3{,}09$, de manera que

$$Y = 0{,}52\,X + 3{,}09$$



c) Para predecir este valor de la variable *B*, utilizamos la recta de regresión, puesto que se supone que describe bien la dependencia media entre ambas variables:

$$Y(X)=Y(2)=4{,}13$$

d) Otra de las fórmulas importantes de la recta de regresión es esta relación por la que nos preguntan, que es:

$$m=\frac{s_{AB}}{s_A^2}=\frac{s_{AB}}{s_A^2}\cdot\frac{s_B}{s_B}=\frac{s_{AB}}{s_A s_B}\cdot\frac{s_B}{s_A}=r\frac{s_B}{s_A},$$

fórmula que relaciona el coeficiente de la recta de regresión de *B* sobre *A* con el coeficiente de correlación lineal entre ambas variables. Esta fórmula indica que son cosas distintas, aunque estén relacionadas. Nótese el efecto del cociente s_B/s_A en general: si es muy cercano a cero puede hacer que la correlación sea grande pero no la pendiente de la recta, mientras que si es muy grande puede hacer que sin apenas correlación la pendiente sea grande.

e) Los residuos se definen como la distancia que hay entre los valores de la variable en la muestra y los valores que predice la recta de regresión, es decir, $r_i=b_i-\hat{b}_i$. Para proceder con orden, construimos la siguiente tabla:

A	-1	-4	0	3	5
B	3	1	2	6	5
Y	2,56	0,99	3,09	4,66	5,71
R	0,44	0,01	-1,09	1,34	-0,71

La desviación típica residual se define como $s_R=\sqrt{s_R^2}=\sqrt{\frac{1}{m}\sum r_i^2}=\sqrt{\frac{1}{m}\sum (b_i-\hat{b}_i)^2}$, pero en vez de utilizar esta fórmula (ahora sí podemos, porque conocemos los valores y_i), vamos a utilizar que

$$\frac{s_R^2}{s_B^2}=1-r^2$$

de donde $s_R^2=(1-r^2)\cdot s_B^2=(1-0{,}89^2)\cdot 3{,}44=0{,}72$, por lo que $s_R=\sqrt{0{,}72}=0{,}85$.

f) La respuesta se obtiene utilizando la siguiente fórmula, que involucra los conceptos que menciona el enunciado: $CV_A = \frac{s_A}{|\bar{a}|}$, de donde se tiene que $|\bar{a}| = 1$, lo que da lugar a dos soluciones:

1) $\bar{a} = 1 \quad \rightarrow \quad \frac{1}{5}(a_1 - 4 + 0 + 3 + 5) = 1 \quad \rightarrow \quad a_1 = 1$

2) $\bar{a} = -1 \quad \rightarrow \quad \frac{1}{5}(a_1 - 4 + 0 + 3 + 5) = -1 \quad \rightarrow \quad a_1 = -9$

Aunque A represente una nota académica, vemos en la muestra original que puede tomar valores negativos (es así por la escala en que se están midiendo esas notas), por lo que ambas soluciones serían válidas.

Quien esté interesado en la Teoría de la Probabilidad, herramienta fundamental de la Estadística para pasar de describir a inferir, puede consultar el archivo:

«Comentarios de Teoría de la Probabilidad»

http://www.Casado-D.org/edu/ComentariosTeoriaProbabilidad.pdf