Introducción a la Estadística Descriptiva a través de casos reales

Paula Lagares Barreiro*
Frederico Perea Rojas-Marcos*
Justo Puerto Albandoz*

MaMaEuSch†
Management Mathematics for European Schools
94342 - CP - 1 - 2001 - 1 - DE - COMENIUS - C21

*University of Seville

†Este proyecto ha sido llevado a cabo con ayuda parical de la Comunidad Europea en el marco del programa Sócrates. El contenido del proyecto no reflejy necesariamente la posición de la Comunidad Europea, ni implica ninguna responsabilidad por su parte.

Índice General

Capítulo 1

Estadística descriptiva unidimensional

Vamos a estudiar una encuesta de opinión. Rellenaréis una encuesta, vamos a ver qué opinais de un montón de temas y a estudiar características como alturas, número de hermanos, etc. Podremos comprobar si vuestras opiniones las comparte más gente y si hay muchos compañeros que tengan algunas características parecidas a las vuestras, por ejemplo ¿cuántos de tus compañeros serán más altos que tú? ¿Y cuántos tienen el mismo número de hermanos que tú? Antes de seguir, plantearemos los principales objetivos que perseguimos en este capítulo.

1.1 Objetivos

- Distinguir los distintos tipos de caracteres estadísticos.
- Determinar qué tipo de tratamiento estadístico conviene hacer, dependiendo de la naturaleza de los datos estudiados.
- Presentar conjuntos de datos con la ayuda de tablas y gráficos.
- Conocer los conceptos de centralización y dispersión de un conjunto de datos.
- Determinar los parámetros de una distribución estadística.
- Estudiar el coeficiente de variación.
- Alentar la curiosidad, a través de la información suministrada en ejercicios y problemas, ante cuestiones de tipo social, ecológico, económico, etc.

1.2 El ejemplo: una encuesta de opinión

A partir de ahora vamos a trabajar con una encuesta de opinión. Queremos saber ciertas cosas sobre los alumnos del mismo curso que vosotros. Os preguntaremos algunos datos y luego nos daréis opinión e información sobre muchos aspectos, como la alimentación, el deporte, etc. Nuestra encuesta será anónima, para que cada uno pueda contestar libremente y sin pensar en que luego los demás sabrán lo que ha contestado. Así, con estos datos, intentaremos plantearnos preguntas interesantes sobre nosotros mismos, que quizás podamos tomar como referente para responder a otras sobre un conjunto más amplio de personas. Por ejemplo

- ¿Cuál es la altura más habitual en tu clase?
- ¿Se puede considerar que tu paga es normal en comparación con la de otros chicos? ¿la mayoría tiene más o menos paga que tú?
- ¿Cuántos hacéis deporte regularmente? ¿Y cuántos desayunan antes de venir?
- ¿Qué coméis más: fruta, leche, legumbres, café, carne, pescado ...?

Pues vamos a ver que analizando las respuestas que tenemos en la encuesta podéis contestar a todas estas preguntas. Seguro que al final del capítulo ya las hemos respondido todas. Pero primero vamos a ir presentando los conceptos que necesitarás para ello.

1.3 Población, individuo y muestra

Antes de comenzar a responder preguntas, tenemos que precisar algunas cosas. ¿Sobre quiénes queremos obtener información? Ya hemos visto que sobre los alumnos de tu curso, luego para nosotros, la *población* no sois sólo vosotros, sino todos los alumnos de tu nivel. Pero nos llevaría mucho tiempo preguntaros a todos, y hemos decidido tomar un grupo representativo de todos los grupos de tu nivel, que en este caso sois vosotros. Así, vosotros sois la *muestra*. Además, a cada elemento de la *población* lo llamaremos *individuo*. Hagamos algunas observaciones sobre lo que acabamos de decir. Lo primero es que nosotros podemos querer estudiar alguna característica en animales, plantas o cosas, por ejemplo, la duración de las baterías en los teléfonos móviles, y en este caso, la población no sería "humana", sino que serían los diferentes modelos de teléfonos móviles. Además, podemos encontrarnos con casos en los que la utilización de muestras esté más justificada aún que en nuestro caso, por diferentes motivos: si queremos conocer lo que votarán los españoles en las próximas elecciones, no podemos preguntarle a todos los españoles mayores de 18 años, porque serían millones de personas y supone mucho dinero y tiempo. Para estudiar, por ejemplo, la durabilidad media de unas determinadas bombillas hasta que se funden, no podemos examinarlas todas, porque cada examen supone que una bombilla se funda, es decir, es un caso en el que el individuo se destruye. Por tanto, en muchas situaciones, el muestreo está justificado por razones económicas, de tiempo o de destrucción con el estudio de los individuos de la población.

Ejercicio 1.3.1 *La Encuesta de Demanda de Estudios Universitarios en Andalucía fue realizada el año 2001 para conocer qué pensaban estudiar y por qué los 65356 estudiantes de 2º de Bachillerato.*

Para ello, se recogieron los datos de 8500 estudiantes de 2º de Bachillerato de toda Andalucía. ¿Podrías decir cuáles son la población y la muestra en este caso? ¿Qué motivos justifican la elección de la muestra?

1.4 Tipos de variables estadísticas: cuantitativas (discretas y continuas) y cualtitativas

Para poder responder correctamente a muchas de nuestras preguntas, lo primero que tenemos que saber es qué tratamiento se le debe dar a los datos. Porque si te fijas, no todos los datos que podemos obtener son del mismo tipo, por ejemplo, pensemos sobre las respuestas a tres de las preguntas de la encuesta

1. La respuesta a la pregunta sexo (hombre o mujer)
2. La respuesta a la pregunta número de hermanos
3. La respuesta a la pregunta altura

Lo primero que podemos observar es que la respuesta a la pregunta 1 **no es numérica** mientras que las de las preguntas dos y tres sí lo son. La característica que corresponde a la respuesta de la pregunta 1 se llama *cualitativa* mientras que las variables correspondientes a las preguntas 2 y 3 se llaman *cuantitativas*. Es claro que las variables cuantitativas permiten realizar cálculos que no podemos hacer con las variables cualitativas. A las distintas posibilidades de la característica se les llama *modalidades* en el caso cualitativo y *valores* en el caso cuantitativo. Vamos a ver ahora qué diferencias podemos encontrar entre las variables 2 y 3, porque es algo más complicada. La variable número de hermanos toma valores numéricos que podríamos llamar "aislados", 0,1,2,3,..., pero no puede tomar cualquier valor entre ellos, es decir, no puede tomar el valor 3.5 por ejemplo. Sin embargo, con la altura no ocurre esto. En realidad, la altura puede tomar cualquier valor dentro de unos límites, podemos medirla con tanta precisión como queramos. Podríamos decir que la variable altura puede tomar todos los valores posibles dentro de un intervalo. Así, a la variable que resulta en el caso 2 se le llama variable *discreta* y a la que resulta del caso 3 se le llama variable *continua*.

Ejercicio 1.4.1 *Indica si las siguientes variables son cualitativas o cuantitativas, y en caso de ser cuantitativas, si son discretas o continuas:*

1. *Número de nacidos en un día*
2. *Grupo sanguíneo de una persona*
3. *Tiempo que se necesita para resolver un problema*
4. *Número de preguntas de un examen*
5. *Temperatura de una persona*
6. *Partido político votado en las últimas elecciones*
7. *Número de goles marcados por un jugador en una temporada*

1.5 Tablas estadísticas: Frecuencias absolutas, relativas y porcentuales. Agrupación de los datos por intervalos

Bueno, es el momento de empezar a manejar los datos que hemos obtenido de la encuesta. Los datos que hemos obtenido de la pregunta número de hermanos son los siguientes: 0 1 3 2 0 1 0 1 1 2 2 3 1 2 1 1 1 1 0 0 4 2 3 1 2 1 2 1 1 0 mientras que los siguientes son los datos que se refieren al peso 52 66 54 70 46 62 59 68 49 50 77 57 63 67 58 54 52 47 74 72 80 82 60 75 53 55 69 67 50 52 Tenemos un montón de curiosidades: ¿cuántos compañeros tienen el mismo número de hermanos que yo?¿Cuántos tienen más?¿y menos? ¿Cuántos pesan más o menos como yo?¿y más?¿y menos? Para contestar a estas preguntas tendríamos que contar cuántas veces aparece cada respuesta. Empecemos por contestar las que se refieren al número de hermanos. Para poder contestar necesitamos saber cuántas veces aparece cada respuesta, así que hacemos recuento:

$$\begin{array}{ll} 0 & ||||| \ | \rightarrow 6 \\ 1 & ||||| \ ||||| \ ||| \rightarrow 13 \\ 2 & ||||| \ || \rightarrow 7 \\ 3 & ||| \rightarrow 3 \\ 4 & | \rightarrow 1 \end{array}$$

Ya sabemos, por ejemplo, que hay 13 personas que tienen 1 hermano. A este número lo llamamos *frecuencia absoluta* y lo denotamos por n_i. ¿Y cuantos tienen como mucho un hermano? Pues los que tengan 0 ó 1 hermanos, es decir $6 + 13 = 19$. A este número lo llamamos *frecuencia absoluta acumulada* en este caso para el valor 1. Denotaremos las frecuencias absolutas acumuladas por N_i Construyamos pues la tabla de frecuencias absolutas y acumuladas

Num hermanos	fr. absoluta	fr. absoluta acumulada
0	6	6
1	13	$13 + 6 = 19$
2	7	$13 + 6 + 7 = 26$
3	3	$13 + 6 + 7 + 3 = 29$
4	1	$13 + 6 + 7 + 3 + 1 = 30$

Es importante que ordenemos los valores de la característica de mayor a menor al representarlos en la tabla, para calcular correctamente las frecuencias acumuladas. Vamos a definir algún otro tipo de frecuencia más, porque es interesante saber qué proporción del total supone cada uno de los números, porque así podremos comparar con otras poblaciones. Por ejemplo, en nuestro caso, 6 alumnos tienen 0 hermanos, pero hemos preguntado en un grupo de 50 personas y sabemos que son 9 personas las que tienen 0 hermanos ¿en cuál de los dos grupos hay una mayor proporción de hijos únicos? Pues es fácil, las proporciones son

$$\frac{6}{30} = 0.2 \quad \text{y} \quad \frac{9}{50} = 0.18$$

luego la proporción es mayor en el grupo de 30 personas. Esta proporción se llama *frecuencia relativa* y se denota por f_i. Si la expresamos en porcentaje (multiplicándola por 100), obtenemos la *frecuencia porcentual*, que en el caso anterior serían el 20% y el 18% respectivamente. Denotaremos por p_i a estas frecuencias. Añadimos ahora estas frecuencias a la tabla que teníamos

Hermanos	fr. absoluta	fr. relativa	fr. porcentual	fr. abs. acum.	fr. rel. acum.
0	6	$\frac{6}{30} = 0.2$	20%	6	0.2
1	13	$\frac{13}{30} = 0.4\widehat{3}$	$43.\widehat{3}\%$	$13 + 6 = 19$	$0.6\widehat{3}$
2	7	$\frac{7}{30} = 0.2\widehat{3}$	$23.\widehat{3}\%$	$13 + 6 + 7 = 26$	$0.8\widehat{6}$
3	3	$\frac{3}{30} = 0.1$	10%	$13 + 6 + 7 + 3 = 29$	$0.9\widehat{6}$
4	1	$\frac{1}{30} = 0.\widehat{3}$	$3.\widehat{3}\%$	$13 + 6 + 7 + 3 + 1 = 30$	1

Veamos ahora el caso de los datos sobre el peso. Recontamos los valores iguales

$$
\begin{array}{ll}
46 & | \to 1 \\
47 & | \to 1 \\
49 & | \to 1 \\
50 & || \to 2 \\
52 & ||| \to 3 \\
53 & | \to 1 \\
54 & || \to 2 \\
55 & | \to 1 \\
57 & | \to 1 \\
58 & | \to 1 \\
59 & | \to 1 \\
60 & | \to 1 \\
62 & | \to 1 \\
63 & | \to 1 \\
66 & | \to 1 \\
67 & || \to 2 \\
68 & | \to 1 \\
69 & | \to 1 \\
70 & | \to 1 \\
72 & | \to 1 \\
74 & | \to 1 \\
75 & | \to 1 \\
77 & | \to 1 \\
80 & | \to 1 \\
82 & | \to 1
\end{array}
$$

Como ves, la mayoría de los valores sólo tienen frecuencia uno, y además, nuestra variable toma 25 valores diferentes, que son demasiados para representar en una misma tabla (y más aún si, como en nuestro caso, tenemos sólo 30 observaciones) ¿Qué podemos hacer para obtener una tabla más representativa de cómo están repartidos los datos? Pues parece lógico agrupar los valores cercanos por intervalos. Sobre la agrupación por intervalos hay toda una teoría que nos habla de la manera correcta de hacer dicha agrupación. Aquí sólo veremos algunas indicaciones importantes:

- El número de clases no debe ser ni muy elevado (entre 6-8 es el número máximo con el que habitualmente trabajamos) ni muy escaso (no tiene sentido agrupar sólo en dos o tres clases, perdemos mucha información)
- Salvo tal vez las dos clases extremas, las clases deben tener la misma amplitud, si no, la información se vería distorsionada.

¿Se te ocurre cuáles pueden ser los intervalos que buscamos? Piénsalo en función del número de clases que quieres obtener, por ejemplo. Observemos lo siguiente: entre el mayor valor (82) y el menor (46) tenemos una diferencia de 36 kg. Si queremos hacer, por ejemplo, 6 clases, pues la amplitud (tamaño del intervalo) debe ser $\frac{36}{6} = 6$. Así, obtendríamos los siguientes intervalos: [46,52], (52,58], (58,64], (64,70], (76,82]. Así obtenemos una posible clasificiación, aunque, por supuesto, puede haber más. En algunos estudios, encontrarás, que la primera clase es del tipo "menor que 52"y la última "mayor que 76". A este tipo de intervalos los consideraremos del mismo tamaño que los anteriores a efectos de cálculo. Una vez decidida la clasificación por intervalos, podemos calcular las frecuencias:

Peso	fr. absoluta	fr. relativa	fr. porcentual	fr. abs. acum.	fr. rel. acum.
[46,52]	8	$0.2\widehat{6}$	$26.\widehat{6}\%$	8	$0.2\widehat{6}$
(52,58]	6	0.2	20%	14	$0.4\widehat{6}$
(58,64]	4	$0.1\widehat{3}$	$13.\widehat{3}\%$	18	0.6
(64,70]	6	0.2	20%	24	0.8
(70,76]	3	0.1	10%	27	0.9
(76,82]	3	0.1	10%	30	1

Además, cuando trabajemos con datos agrupados en intervalos, necesitaremos escoger un representante de cada uno de los intervalos, lo que llamaremos *marca de clase*, que será el punto medio del intervalo en cuestión (extremo inferior del intervalo más el extremo superior del intervalo, dividido entre 2).

Ejercicio 1.5.1 *Calcula la tabla de frecuencias de las respuestas a la pregunta 1.3 de la encuesta y de las respuestas a la pregunta "altura", decidiendo previamente si es necesaria una agrupación por intervalos de los datos o no.*

1.6 Representaciones gráficas

Una vez que has calculado las tablas de frecuencias, tu profesor te pide que expongas ante el resto de tus compañeros las conclusiones que has obtenido. Podrías presentar las tablas de frecuencias y hablar sobre las conclusiones más relevantes, pero ¿hay alguna forma de presentar los datos de manera que las principales características de estos sean visibles de una manera sencilla? Obviamente, la respuesta es que sí. Como habrás observado, tanto en libros, como, fundamentalmente, en los medios de comunicación, los datos suelen presentarse a través de gráficos, que resultan más atractivos a la vista que una tabla de frecuencias, además de que permiten una interpretación más sencilla de los datos de los que disponemos. En esta sección vamos a intentar conocer la mayoría de los gráficos, y vamos a hacer especial hincapié en lo importante que es elegir el tipo adecuado de gráfico según los datos con los que trabajemos. Ya que tenemos las tablas de número de hermanos y del peso, los utilizaremos para ir introduciendo los diferentes tipos de gráficos.

1.6.1 Diagrama de barras

El primer tipo de gráfico que veremos es el **diagrama de barras**. Este es un gráfico que se usa

tanto para variables cualitativas como para variables discretas no agrupadas por intervalos. Como sabemos que nuestros datos sobre número de hermanos corresponden a una variable discreta, vamos a ver cómo se construye un diagrama de barras utilizando esos datos. En el eje de abscisas (el eje OX) colocamos las *modalidades* si la variable es cualitativa o los valores en caso de que la variable sea discreta, en nuestro caso, los valores 0, 1, 2, 3 y 4. Sobre cada uno de estos valores se levanta una barra (o rectángulo) de igual base (que no se solapen entre ellos), cuya altura sea proporcional a la frecuencia. En nuestro caso, quedaría más o menos de la siguiente manera: En ocasiones, este

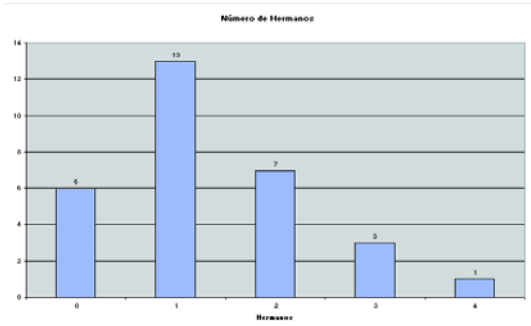

Figura 1.1: hermanos (barras vericales)

tipo de gráfica también se presenta con las barras en horizontal, de la siguiente manera

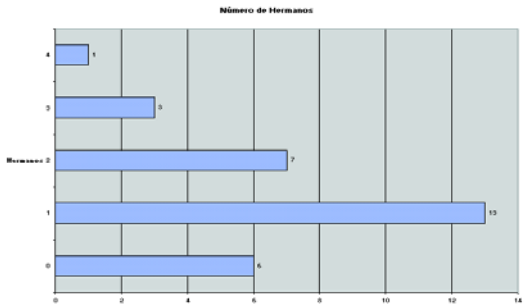

Figura 1.2: hermanos (barras horizontales)

1.6.2 Histogramas

El **histograma** es un gráfico muy similar al anterior, pero es el que utilizaremos para variables agrupadas por intervalos. Nosotros construiremos un histograma para la variable peso. Se realiza, como el anterior, sobre ejes cartesianos, representando en el eje OX los intervalos y levantando rectángulos que tienen como base la amplitud de los distintos intervalos y una altura tal que el

área del rectángulo sea proporcional a la frecuencia correspondiente al intervalo. En este tipo de gráfico son muy importantes las áreas de los rectángulos, porque no representamos una barra correspondiente a un punto, sino que el ancho de la barra representa a nuestro intervalo. Así, si los intervalos son de la misma amplitud, la altura suele corresponder a la frecuencia, pero si no es así, hay que modificar la altura para mantener la proporción entre la frecuencia y el área. Nuestro histograma sobre la variable peso, que tenemos agrupada del ejemplo anterior, podría tener el siguiente aspecto También podemos representarlo con los rectángulos en horizontal, así tendríamos:

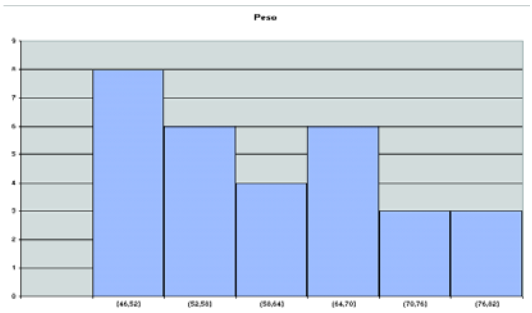

Figura 1.3: peso (histograma)

Seguro que alguna vez has visto una pirámide de población en algún medio de comunicación o

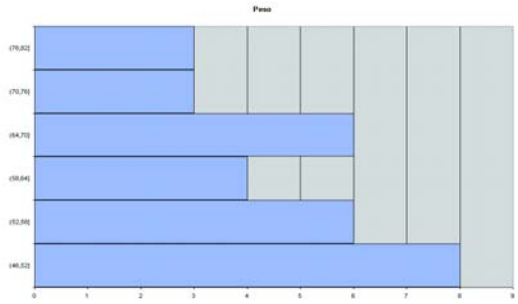

Figura 1.4: peso (histograma)

publicación. Pues una pirámide de población no es más que dos histogramas horizontales (uno para las mujeres y otro para los hombres, en los que se representa el número de habitantes agrupados por edad. (añadir gráfico)

1.6.3 Polígonos de frecuencias

El siguiente tipo de gráfico que veremos son los **polígonos de frecuencias**. Este gráfico se utiliza

para el caso de variables cuantitativas, tanto discretas como continuas. Para realizarlos, partimos del diagrama de barras o del histograma, según la variable sea agrupada o no agrupada. Lo que debemos hacer es unir mediante una línea los puntos medios de las bases superiores del diagrama de barras o del histograma, según corresponda. En nuestros dos ejemplos tendríamos, que para el caso del número de hermanos obtenemos El caso del peso es también algo diferente. En este gráfico, el

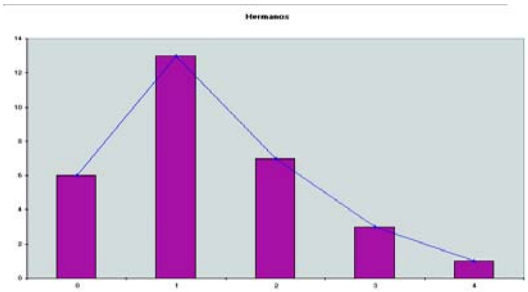

Figura 1.5: hermanos (polígono de frecuencias)

área por debajo de la línea representa los datos que tenemos, al igual que en el histograma, puesto que estamos hablando de la amplitud completa del intervalo. El gráfico quedaría como sigue Todos

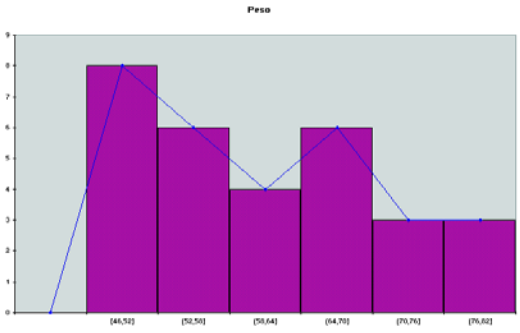

Figura 1.6: peso (polígono de frecuencias)

los gráficos que hemos visto anteriormente los podemos representar no sólo para las frecuencias absolutas, sino también para las relativas y para las acumuladas.

1.6.4 Diagrama de sectores

El siguiente tipo de gráfico que vamos a ver seguro que lo conoces, se llama **diagrama de sectores o de tarta**. En él, a cada modalidad o valor se le asigna un sector circular de área

proporcional a la figura que representan. Este gráfico se utiliza para variables cualitativas o para variables discretas sin agrupar.

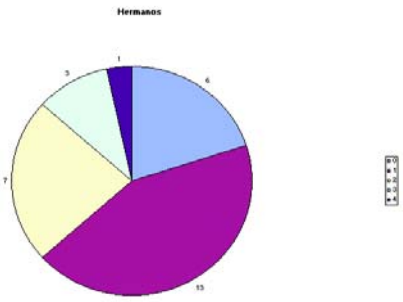

Figura 1.7: hermanos (diagrama sectores)

1.6.5 Pictogramas

Ahora veremos otro gráfico que también es muy frecuente en prensa, se denominan **pictogramas**. Son gráficos con dibujos alusivos al carácter que se está estudiando y cuyo tamaño (atención: no sólo la altura, sino el tamaño, en término de área) es proporcional a la frecuencia que representan, frecuencia que suele indicarse junto al dibujo para evitar confusiones. (añadir ejemplo)

1.6.6 Diagrama de tallo y hojas

Existe otro tipo de gráfico que está entre el recuento de casos y el gráfico, y que se llama **diagrama de tallo y hojas**. Vamos a ir viendo cómo se construye con el ejemplo del peso. Recordamos que los datos que tenemos son 52 66 54 70 46 62 59 68 49 50 77 57 63 67 58 54 52 47 74 72 80 82 60 75 53 55 69 67 50 52 El diagrama de tallo y hojas, lo primero que hace es indicar, en una columna las diferentes cifras correspondientes a las decenas que podemos encontrar en el conjunto de datos, en nuestro caso, como los valores oscilan entre 46 y 82, tendremos que poner 4, 5, 6, 7 y 8, de la siguiente manera

4 |
5 |
6 |
7 |
8 |

A continuación tomamos la primera observación, 52, y colocamos la cifra de las unidades al lado de la que corresponde a su decena, es decir

4	
5	2
6	
7	
8	

Así, seguimos colocando la cifra de las unidades al lado de la correspondiente cifra de las decenas para el resto de los valores. Obtenemos algo como:

4	697
5	249078423502
6	62837097
7	07425
8	02

Como ves, obtenemos algo similar (que no igual) a un diagrama de barras o un histograma horizontal. Obviamente, también lo podríamos haber hecho en vertical, y nos quedaría algo como:

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

que como ves, se parece a un histograma o un diagrama de barras habitual aunque *no lo es*. Pero el diagrama de tallo y hojas nos sirve para darnos orientación sobre cómo se distribuyen nuestros datos. En realidad nosotros hemos dividido por decenas (de 40 a 49, de 50 a 59, ...), pero podríamos hacer una división también en grupos de 5 (de 40 a 44, de 45 a 49, de 50 a 54, ...), sin más que poner dos veces cada una de las decenas, a continuación de la primera los valores de las unidades que estén entre 0 y 4, y a continuación de la segunda los valores que están entre 5 y 9. Para el caso horizontal tendríamos

4	
4	697
5	24042302
5	9785
6	230
6	68797
7	042
7	75
8	02
8	

1.6.7 Algunas observaciones

Por ejemplo, imagina que te damos los dos gráficos siguientes referidos a los beneficios de una empresa: De las dos ¿cuál preferirías que fuera tu empresa? Seguro que casi todos estais de acuerdo

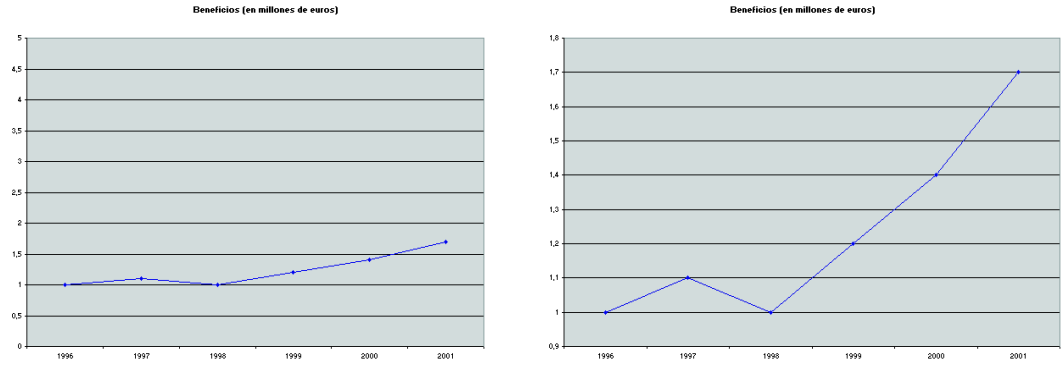

Figura 1.8: beneficios (empresa 1 y empresa 2)

en que la 2 es mejor que la 1, porque parece que tiene más beneficios, pero la realidad es que los datos de las dos gráficas son los mismos. Tan sólo hemos cambiado la escala del eje OY. *¿Ejemplos reales?*. Haremos, antes de pasar a la siguiente sección, algunas reflexiones. Los gráficos son una herramienta muy útil y que permiten una fácil interpretación de los datos que se manejan pero es necesario realizarlos correctamente para que dicha interpretación no nos induzca a error. Es muy importante mantener las proporciones en las figuras que se representan, así como asegurar que las escalas de los ejes se mantienen también proporcionales puesto que si hacemos cambios en estos sentidos los gráficos tienen diferente apariencia y pueden ser mal interpretados.

Ejercicio 1.6.1 *Realiza varios tipos de gráficos, utilizando los datos de las variables altura, edad y la respuesta a la pregunta 2.4.*

1.7 Medidas de centralización: media, mediana, moda, cuantiles

Supongamos ahora que nos vamos de viaje de fin de curso y queremos ganar algún dinerito, así que hemos decidido que vamos a vender camisetas, pero no sabemos a qué precio. Lo único que sabemos es que el fabricante nos las vende a 4 euros y nos gustaría sacar beneficios pero sin abusar. Nos parece que la paga semanal es una buena referencia para saber cuánto podría gastarse la mayoría de la gente. Así que utilizaremos los valores que tenemos de las pagas semanales 6 8 10 5 15 20 9 10 9 9 20 15 12 6 15 12 10 25 20 30 15 12 9 20 6 9 10 25 9 9 Tenemos los 30 valores, pero nosotros necesitamos un único valor que represente a todos. ¿Qué valor podemos elegir? Una

buena solución sería elegir un valor medio de todos los que tenemos, para ello, los sumamos todos y los dividimos entre el número total de valores, y obtendríamos

$$\overline{x} = \frac{6+8+10+5+15+20+9+10+9+9+20+15+12+6+15+12+10+25}{30}+$$

$$+\frac{20+30+15+12+9+20+6+9+10+25+9+9}{30} = \frac{390}{30} = 13$$

Ya hemos obtenido una primera cantidad como posible precio, 13 euros. A esta cantidad que acabamos de calcular, la llamamos *media aritmética*. Pero otra posibilidad sería elegir como representante de todos los valores, el valor que aparece más frecuentemente. En nuestro caso, el valor más frecuente es el 9, que también podríamos utilizarlo como posible precio. A la cantidad más frecuente la llamamos *moda*. Pero ninguna de las dos cantidades anteriores nos dan información sobre el número de personas que podría pagar la camiseta. Así que se nos ocurre otra opción. Vamos a ordenar los datos que tenemos 5 6 6 6 8 9 9 9 9 9 9 9 10 10 10 10 12 12 12 15 15 15 15 20 20 20 20 25 25 30 Entonces, ahora queremos encontrar el valor que deje la mitad de los valores a cada lado. Los valores que ocupan el lugar 15 y 16 dejan 14 valores a cada lado, como ambos son el 10, podemos considerar que es el 10 el valor que deja el 50% de los valores a cada lado y podemos considerarlo como posible cantidad. A esta cantidad la llamaremos *mediana*. Igual que hemos podido pensar en una cantidad que puedan pagar la mitad, podemos decidir que puedan pagarlo el 75% de la población, es decir, encontrar una cantidad que deje el 25% a la izquierda (es decir, sólo el 25% de los datos sería menor), o cualquier otro porcentaje. A estas cantidades las llamamos *cuantiles*. De las tres cantidades obtenidas, podemos elegir la que más se ajuste a nuestra situación. No siempre las tres serán válidas, pero son tres medidas que nos dan una idea de dónde está el centro de nuestros datos. Son las principales *medidas de centralización*. Vamos a ver ahora la definición rigurosa de los conceptos que acabamos de presentar. Nos referiremos sólo a variables. Suponemos que se ha observado una variable en n individuos y se han obtenido k valores diferentes $x_1, x_2, \dots x_k$, cada uno con una frecuencia absoluta de $n_1, n_2, \dots n_k$ donde n_i es la frecuencia absoluta del valor x_i. Por $N_i = \sum_{j \leq i} n_j$ denotamos a la frecuencia absoluta acumulada del valor x_i y por $f_i = \frac{n_i}{n}$ a la frecuencia relativa de x_i. Si los valores observados en los n individuos, se agrupan en intervalos, podemos suponer que se toman h intervalos que notaremos

$$(L_0, L_1], (L_1, L_2], \dots (L_{h-1}, L_h]$$

cuyas marcas de clase serán $c_1, c_2, \dots c_h$. A las frecuencias absolutas asociadas las denotaremos por $n_1, n_2, \dots, n_h$, a las frecuencias absolutas acumuladas por $N_1, N_2, \dots, N_h = n$ y a las frecuencias relativas por $f_1, f_2, \dots, f_h$ La *media aritmética* o, simplemente *media* se calcula sumando todos los elementos y dividiendo por el número total de elementos de la población, es decir

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}$$

La media tiene las siguientes características:

- Es el centro de gravedad de la distribución y es única para cada distribución.
- Cuando aparecen valores extremos y poco significativos (demasiado grandes o demasiado pequeños), la media puede dejar de ser representativa.

- No tiene sentido en el caso de una variable cualitativa ni cuando existen datos agrupados con algún intervalo no acotado.
- Para variables agrupadas, los x_i serán las marcas de clase de cada intervalo.

Además, la media cumple las siguientes propiedades:

- Si se suma una constante a todos los valores, la media aumenta en dicha constante.
- Si se multiplican todos los valores de la variable por una constante, la media queda multiplicada por dicha constante.

La *moda* se suele definir como el valor más frecuente. En el caso de una variable no agrupada, es el valor de la variable que más se repite. En el caso de una variable agrupada por intervalos de igual amplitud se busca el intervalo de mayor frecuencia (intervalo o clase modal) y se aproxima la moda por el valor obtenido al aplicar la fórmula

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

donde:
L_{i-1} es el límite inferior del intervalo modal.
n_i es la frecuencia absoluta del intervalo modal.
n_{i-1} es la frecuencia absoluta del intervalo anterior al intervalo modal.
n_{i+1} es la frecuencia absoluta del intervalo posterior al intervalo modal.
c_i es la amplitud del intervalo.

La moda cumple que

- Puede ser que exista más de una moda. En dicho caso, se dice que la distribución es bimodal, trimodal, ..., según el número de valores que presentan la mayor frecuencia absoluta.
- La moda es menos representativa que la media, a excepción de las distribuciones con datos cualitativos.
- Si los intervalos no tienen la misma amplitud, se busca el intervalo de mayor densidad de frecuencia (que es el cociente entre la frecuencia absoluta y la amplitud del intervalo: $\frac{n_i}{c_i}$) y se calcula con la fórmula anterior.

La *mediana* es, en el caso de una variable no agrupada, y una vez ordenados los datos, el valor central si el número de observaciones es impar y la media de los valores centrales si es par. En el caso de una variable agrupada, hemos de buscar el intervalo central (en el que se encuentre el o los valores centrales), es decir, áquel en el que N_i supera por primera vez $\frac{n}{2}$, $N_{i-1} \leq N_i$, y aplicar la fórmula

$$Me = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot c_i$$

donde:
L_{i-1} es el límite inferior del intervalo.
n_i es la frecuencia absoluta del intervalo.
N_{i-1} es la frecuencia absoluta acumulada del intervalo anterior.
n es el número de datos.
c_i es la amplitud del intervalo.

Además, los *cuantiles* son medidas de posición que generalizan el concepto de mediana. Definiremos ahora los *centiles o percentiles*,

los *cuartiles* y los *deciles*. Suponemos que tenemos los datos ordenados. Los *centiles* o *percentiles* son los valores de la variable que dejan a su izquierda un determinado porcentaje de la población. Se representan por C_h o P_h, donde h indica el porcentaje, $h = 1, 2, \ldots, 99$. En el caso de una variable agrupada, una vez obtenido el intervalo en que se encuentra el centil, se aplica la siguiente fórmula

$$P_h = C_h = L_{i-1} + \frac{h \cdot \frac{n}{100} - N_{i-1}}{n_i} \cdot c_i$$

donde cada elemento tiene el mismo significado que en el cálculo de la mediana. Los *cuartiles* son los valores que, una vez ordenados los datos, dividen a la variable en 4 grupos iguales. En cada uno de ellos hay un 25% de individuos de la población o muestra. Se representan por Q_1, Q_2 y Q_3 y verifican $Q_1 = C_{25}$, $Q_2 = C_{50} = Me$, $Q_3 = C_{75}$. Los *deciles* son los valores que, una vez ordenados los datos, dividen a la misma en 10 partes iguales, de modo que entre 2 deciles hay un 10% de los individuos de la población o muestra. Se representan por $D_1, D_2, D_3, \ldots, D_9$. Verifican $D_1 = C_{10}, D_2 = C_{20}, D_3 = C_{30}, \ldots D_9 = C_{90}$.

Ejercicio 1.7.1 *Para los datos de número de hermanos y de peso, calcular media, moda y mediana, y los cuantiles:* $Q_1, Q_3, C_{30}, C_{74}, D_4, D_9$

1.8 Medidas de dispersión: Rango, varianza, desviación típica

Imagina que tenemos 3 conjuntos de personas y nos dicen que en todos los casos, la media del peso es 55. ¿Significa esto que los tres conjuntos de datos son iguales o similares? Conseguimos los datos originales y nos encontramos con que las observaciones son las siguientes:

Grupo 1: 55 55 55 55 55 55 55

Grupo 2: 47 51 54 55 56 59 63 vemos que, aunque la media es la misma, los conjuntos de datos

Grupo 3: 39 47 53 55 57 63 71

son muy diferentes. Fíjate si hacemos el diagrama de tallo y hojas lo que obtenemos

		5																	
		5																	
		5						9											
		5						6											
		5						5						7					
		5						4						5					
		5					7	1	3			9	7	1	3	1			
3	4	5	6	7		3	4	5	6	7		3	4	5	6	7			

Entonces ¿cómo podemos detectar esas diferencias entre los conjuntos de datos? Parece que las medidas de centralización no nos proporcionan información suficiente en muchas situaciones, así que debemos encontrar alguna otra cantidad que nos diga cómo de lejos están los datos entre ellos y de la media, es decir, nos surje la necesidad de medir la dispersión de los datos. Lo primero que vemos es que en el primer caso todos los datos son iguales, en el segundo hay más diferencia entre el mayor y el menor, y en el tercero más aún que en el segundo. Exactamente tenemos que

$55 - 55 = 0$
$63 - 47 = 16$ A esta cantidad la llamamos *rango* de los datos. Sin embargo, aunque es muy fácil
$71 - 39 = 32$
de calcular, no se usa demasiado, porque si hay un sólo valor muy grande o muy pequeño, el rango varía mucho, así que no siempre es una medida útil. ¿Cómo podríamos encontrar un número que nos dé una aproximación de la distancia de los datos a la media? Pues podemos calcular todas las diferencias (en valor absoluto) entre las observaciones y la media y luego calcular la media de esas diferencias. A esta cantidad la llamamos *desviación media*. Calculemos la desviación media del grupo 2 de datos, tenemos

$$\frac{|47-55|+|51-55|+|54-55|+|55-55|+|56-55|+|59-55|+|63-55|}{7} =$$

$$= \frac{8+4+1+0+1+4+8}{7} = \frac{26}{7} = 3.714$$

Sin embargo, habitualmente se usa otra medida de la variabilidad, que responde a la media de los cuadrados de las desviaciones de los datos respecto a la media, así conseguimos que las desviaciones mayores influyan más que las pequeñas. Pero vamos a ver la definición rigurosa de todos estos conceptos. El *rango o recorrido* es la diferencia entre el valor mayor y el menor de la variable si ésta es no agrupada. Si la variable es agrupada, se calcula la diferencia entre el límite superior del último intervalo y el límite inferior del primer intervalo. El valor del rango sólo tiene en cuenta el mayor y el menor elemento, en su valor no influyen los demás elementos de la distribución. Por ejemplo, los siguientes podrían ser dos conjuntos de datos representados en una recta para ambos tendríamos el

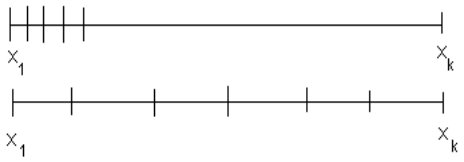

Figura 1.9: rango

mismo rango, ya que la diferencia entre x_k y x_1 es la misma, pero está claro que los dos conjuntos de datos son muy diferentes. El *rango intercuartílico* es la diferencia entre el primer y el tercer cuartil, y nos da una franja entre la que se encuentra el 50% de la población. La *desviación media* es la media de las desviaciones de los valores de la variable respecto a la media de la distribución. Se llama desviación respecto de la media al valor absoluto de la diferencia de los valores entre la variable y la media ($|x_i - \overline{x}|$), luego la expresión de la desviación media es

$$DM = \frac{\sum_{i=1}^{k} |x_i - \overline{x}| \cdot n_i}{n}$$

es una medida muy poco utilizada por lo complicado de su cálculo, ya que hay que tratar con la función valor absoluto. Si la desviación media es muy pequeña, indica que hay una gran concentración

de valores en torno a la media. Existe también, aunque se utiliza menos, la *desviación respecto a la mediana*, que es la media de las desviaciones con respecto a la mediana

$$D = \frac{\sum_{i=1}^{k} |x_i - Me| \cdot n_i}{n}$$

La *varianza* es la media de los cuadrados de las desviaciones respecto a la media. Se representa por S^2 y su expresión es

$$S^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 \cdot n_i}{n} = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{n} - \overline{x}^2$$

La varianza cumple que

- Al tomar los cuadrados de las desviaciones tiene la ventaja de que las desviaciones grandes afectan más al resultado.
- Las unidades de S^2 no son las mismas que las de la muestra, ya que estamos elevando las desviaciones al cuadrado.
- La varianza es siempre positiva. Es nula cuando todos los valores coinciden con la media.

Definimos la *cuasivarianza* como

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 \cdot n_i}{n-1}$$

cuya relación con la varianza es $S^2 = \frac{n-1}{n} s^2$. Esta medida será muy útil más adelante cuando veamos la inferencia estadística. En ocasiones, también se denota por S_c^2 La *desviación típica* es la raíz cuadrada de la varianza. Se representa por S y su expresión es

$$S = +\sqrt{\frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 \cdot n_i}{n}} = +\sqrt{\frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{n} - \overline{x}^2} = +\sqrt{\overline{x^2} - \overline{x}^2}$$

La desviación típica tiene como características fundamentales

- Es la unidad de dispersión más utilizada.
- Las unidades de S son las mismas que las de la muestra.
- La desviación típica es siempre positiva o cero.

Además, varianza y desviación típica verifican que:

- Si a los valores de una variable se les suma la misma constante, la varianza y la desviación típica no varían.
- Si a los valores de una variable se les multiplica por la misma constante positiva, la varianza queda multiplicada por el cuadrado de la constante y la desviación típica queda multiplicada por dicha constante

1.9 Utilización conjunta de la media y la desviación típica: el teorema de Tchebicheff, el coeficiente de variación de Pearson, tipificación de variables

1.9.1 El teorema de Tchebicheff

Ya hemos encontrado las cantidades que nos dan el centro de los datos y que miden cómo de dispersos están, pero seguimos necesitando más información. Recordemos los datos sobre el número de hermanos:

Núm hermanos	fr. absoluta
0	6
1	13
2	7
3	3
4	1

entonces tenemos que

$$\overline{x} = 1.33333, \quad S^2 = 1.022, \quad S = 1.011$$

¿Cuánta gente hay alrededor de la media?¿Hay muchos compañeros que tengan entre 1 y 2 hermanos? Tomemos un intervalo alrededor de la media, de la forma $(\overline{x}-a, \overline{x}+a)$. Ya que la varianza y la desviación típica miden la dispersión ¿por qué no las utilizamos? ¿cuál usarías de las dos? Bueno, en principio deberíamos descartar la varianza, porque no debemos sumarla con la media, dado que no tienen las mismas unidades. Tomemos entonces la desviación típica, es decir, tomemos $a = S$. Entonces obtenemos el intervalo $(1.3333 - 1.011, 1.3333 + 1.011) = (0.3223, 2.3443)$. Dentro de este intervalo están las personas que tienen 1 y 2 hermanos, que son 20 de los 30 alumnos, es decir, el 66% de los alumnos. ¿Y si en vez de restar y sumar S lo hacemos para $2S$? Obtenemos el intervalo $(1.3333 - 2.022, 1.3333 + 2.022) = (-0.6887, 3.3553)$. En este intervalo ya tenemos 29 de los 30 datos, es decir un 96.6%. Obviamente, si restamos y sumamos $3S$, en el intervalo que obtenemos ya están todos los datos. Pero ¿esto ocurre siempre? ¿se concentran siempre tantos datos en esos intervalos? Vamos a ver otro ejemplo, el de las pagas: Tenemos

$$\overline{x} = 13, \quad S^2 = 39.2, \quad S = 6.26$$

Entonces
$(13 - 6.26, 13 + 6.26) = (6.74, 19.26) \rightarrow$ contiene 19 datos (63%)
$(13 - 12.52, 13 + 12.52) = (0.48, 25.52) \rightarrow$ contiene 29 datos (96%)
$(13 - 18.78, 13 + 18.78) = (-5.78, 31.78) \rightarrow$ contiene 30 datos (100%)

Como ves, tenemos unos resultados muy similares. Esto es porque hay un teorema que asegura que en estos intervalos hay al menos un determinado porcentaje de los datos, exactamente dice que en un intervalo de la forma $(\overline{x} - aS, \overline{x} - aS)$ con $a > 1$ hay <u>al menos un</u> $100(1 - \frac{1}{a^2})\%$ de los datos. Este resultado se conoce como teorema de Tchebicheff.

1.9.2 El coeficiente de variación de Pearson

Imagínate que trabajamos ahora con los datos de peso y altura. Tenemos que para el peso

$$\overline{x} = 60.8, \quad S^2 = 99.56, \quad S = 9.97$$

mientras que para la altura tenemos

$$\overline{x} = 1.7133, \quad S^2 = 0.0128, \quad S = 0.1132$$

ahora ¿en cuál de los dos casos hay más variabilidad? Se nos podría ocurrir pensar que en el peso, porque la varianza y la desviación típica son mayores, pero mira lo que ocurre si hacemos los mismos cálculos con la altura en centímetros:

$$\overline{x} = 171.33, \quad S^2 = 128.35, \quad S = 11.32$$

Si ahora nos hacemos la pregunta de nuevo ¿qué podemos contestar? La realidad es que no podemos comparar las desviaciones típicas ni las varianzas porque dependen de las unidades, igual que la media. Debemos encontrar una cantidad que no tenga unidades. De momento, sólo sabemos que la desviación típica y la media tienen las mismas unidades así que ¿cómo podemos conseguir una cantidad adimensional? Pues podemos dividirlas, así obtenemos lo que se conoce como *coeficiente de variación de Pearson*

$$CV = \frac{S}{\overline{x}}$$

Si lo calculamos para nuestros dos casos, tenemos, para el peso

$$CV = \frac{9.97}{60.8} = 0.163$$

mientras que para la altura

$$CV = \frac{11.32}{171.33} = \frac{0.1132}{1.7133} = 0.066$$

luego el peso presenta más dispersión que la altura.

1.9.3 Tipificación de variables

Pero aún pueden pasar más cosas. Imagina que mides 1.74 y tienes una amiga en la clase de al lado que mide igual que tú. Pero dentro de cada clase ¿cuál de las dos es más alta? ¿cómo podemos compararla, si sólo sabemos que en la clase de tu amiga la media es 1.708 y la desviación típica 12.53? Existe una manera de transformar estos valores en cantidades "comparables". Este método se llama *tipificación* y consiste en restarle la media a la observación y dividir la cantidad obtenida entre la desviación típica. Con esto conseguimos, si lo hacemos para todas las observaciones, la media de los nuevos valores sea 0 y la desviación típica 1, y así serían observaciones comparables. Para nuestro ejemplo, los dos valores tipificados serían

$$z_1 = \frac{1.74 - 1.7133}{0.1132} = 0.235$$

$$z_2 = \frac{1.74 - 1.708}{0.1253} = 0.255$$

Luego llegamos a la conclusión de que la amiga es más alta (dentro de su clase) puesto que el valor tipificado que corresponde a su observación es mayor. La expresión del valor tipificado correspondiente a una observación x_i es

$$z_i = \frac{x_i - \overline{x}}{S}$$

Capítulo 2

Estadística Descriptiva Bidimensional

En el capítulo anterior estuvimos trabajando con los datos que obtuvimos de la encuesta, obteniendo las primeras conclusiones. Pero no vamos a conformarnos con lo que ya hemos obtenido, porque de esa mismos datos podemos obtener más información con algunas técnicas que veremos a continuación. Antes de continuar, los objetivos en este capítulo son los siguientes.

2.1 Objetivos

- Representar e interpretar un conjunto de valores de dos variables mediante una nube de puntos
- Identificar un conjunto de valores de dos variables dados en forma de tabla o nube de puntos como una distribución bidimensional.
- Interpretar la relación entre dos variables a partir de la nube de puntos, determinando de forma intuitiva si es positiva o negativa, si es funcional o no y, en este caso, si se aproxima a una recta.
- Comparar los aspectos globales de varias distribuciones mediante su nube de puntos.
- Asignar nubes de puntos dadas a diferentes tipos de fenómenos.
- Determinar la relación entre las medias de cada una de las variables con la nube de puntos.
- Encontrar, de forma gráfica, una recta que se ajuste a la nube de puntos.
- Estimar el coeficiente de correlación a partir de una nube de puntos.
- Analizar el grado de relación entre las dos variables, conociendo el coeficiente de correlación.

- Calcular el coeficiente de correlación de distribuciones bidimensionales y hallar las rectas de regresión.
- Hacer predicciones a partir de la recta de regresión.

2.2 El ejemplo: una encuesta de opinión

A lo largo de este capítulo seguiremos profundizando en el análisis de la encuesta de opinión con la que ya comenzamos a trabajar. A partir de la información que ya tenemos, procuraremos responder a preguntas del tipo:

- ¿Hay relación entre la paga que recibís y el número de hermanos que tenéis?
- ¿Influye el deporte que hacéis sobre cuánto fumáis o cuánto bebéis?
- ¿Podemos medir exactamente estas relaciones?

A lo largo del capítulo pretendemos contestar a estas preguntas y a otras diferentes. Iremos presentando los conceptos necesarios para ellos a partir de ahora.

2.3 Introducción y tablas simples

Podríamos pensar en un montón de variables que pueden influir unas sobre otras. Por ejemplo, se nos puede ocurrir pensar que cuanto mayores sois más paga tenéis. Vamos a intentar ver si eso es cierto, así que como ya sabéis del capítulo anterior, para poder obtener alguna conclusión, lo primero que debemos hacer es organizar los datos. Recordamos que los datos de edades y de pagas que tenemos son los siguientes:

Edad	Paga	Edad	Paga
16	6	17	12
16	8	16	10
16	10	18	25
16	5	18	20
17	15	18	30
18	20	19	15
16	9	17	12
17	10	16	9
17	9	19	20
17	9	16	6
19	20	16	9
16	15	16	10
17	12	17	25
16	6	16	9
17	15	16	9

Estos son los *pares* de datos que hemos obtenido. Comencemos agrupando los pares que son iguales. Obtenemos lo siguiente

Edad	Paga	Núm. Personas
16	5	1
16	6	3
16	8	1
16	9	5
16	10	3
16	15	1
17	9	2
17	10	1
17	12	3
17	15	2
17	25	1
18	20	2
18	25	1
18	30	1
19	15	1
19	20	2

A esta tabla que acabamos de construir la llamaremos tabla simple y será el punto de partida para nuestro análisis.

2.4 Tablas de frecuencias, distribuciones marginales y condicionadas

¿Te resulta sencillo obtener conclusiones de la tabla anterior? ¿Podemos encontrar alguna manera alternativa de representar los datos? La idea es evitar las repeticiones (aparecen en la primera columna muchas veces repetida cada edad y en la segunda el valor de las pagas). Agrupamos los datos de la siguiente manera

	Edad			
Paga	16	17	18	19
5	1			
6	3			
8	1			
9	5	2		
10	3	1		
12		3		
15	1	2		1
20			2	2
25		1	1	
30			1	

Esta tabla nos permite una visión más global del reparto de las frecuencias y es más útil cuanto más pares de valores diferentes tenemos. La llamamos *tabla de doble entrada* cuando lo que reprsentamos son variables y *tabla de contingencia* cuando estudiamos dos caracteres cualitativos. Pero de esta tabla ¿podemos obtener el total de personas cuya paga es 12 euros? ¿y el número de personas que tienen 17 años? La respuesta es, obviamente que sí. Observa que todas puedes sumar todas las frecuencias que aparecen en la fila correspondiente al 12 y así obtendrías el número de personas cuya paga es 12. Análogamente puedes obtener el número de personas que tienen 17 años sumando las frecuencias correspondientes a la columna encabezada por el 17. Añadimos estas cantidades a nuestra tabla

	Edad				
Paga	16	17	18	19	Tot
5	1				1
6	3				3
8	1				1
9	5	2			7
10	3	1			4
12		3			3
15	1	2		1	4
20			2	2	4
25		1	1		2
30			1		1
Tot	14	9	4	3	30

En realidad, lo que estás obteniendo son los valores de cada una de las variables idependientemente de la otra. A estos valores los llamamos *distribuciones marginales* de las variables estadísticas. Para tener la distribución marginal completa de la variable edad tomamos la primera y la última fila

Edad	16	17	18	19
frecuencias	14	9	4	3

Igualmente para la variable paga tomamos los datos de la primera y última columnas.

Ejercicio 2.4.1 *¿Podrías construir una tabla similar a la anterior para la variable paga?*

De manera genérica, una tabla de doble entrada es de la siguiente forma:

	Y						
X	y_1	y_2	$\dots$	y_p	$\dots$	y_m	Tot
x_1	n_{11}	n_{12}	$\dots$	n_{1p}	$\dots$	n_{1m}	n_{1*}
x_2	n_{21}	n_{22}	$\dots$	n_{2p}	$\dots$	n_{2m}	n_{2*}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_s	n_{s1}	n_{s2}	$\dots$	n_{sp}	$\dots$	n_{sm}	n_{s*}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kp}	$\dots$	n_{km}	n_{k*}
Tot	n_{*1}	n_{*2}	$\dots$	n_{*p}	$\dots$	n_{*m}	n

donde los valores o modalidades de X son $x_1, x_2, \ldots, x_k$ y los de Y son $y_1, y_2, \ldots, y_m$; n_{ij} indica el número de individuos que presentan la modalidad x_i de la variable X y la modalidad y_j de la variable Y. Asimismo, n_{i*} indica el número de individuos que presentan la modalidad x_i y n_{*j} el número de individuos que presentan la modalidad y_j. n es el número total de individuos de la población o muestra.

Una vez que conocemos las distribuciones marginales podemos calcular la media y la desviación típica de cada una de ellas tratándolas como variables unidimensionales. Su expresión sería:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_{i*}}{n} \qquad S_x = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \overline{x}) n_{i*}}{n}}$$

$$\overline{y} = \frac{\sum_{j=1}^{m} y_j n_{*j}}{n} \qquad S_y = \sqrt{\frac{\sum_{j=1}^{m} (y_j - \overline{y}) n_{*j}}{n}}$$

Ejercicio 2.4.2 *¿Cuáles son la media y la desviación típica de las variables edad y paga?*

Uno de tus comañeros tiene una curiosidad. Él tiene 17 años y quiere saber si su paga está entre las mayores o las menores para, en caso de que sea de las menores, pedirle una subida a su padre. Para eso se quiere comparar con los compañeros que tienen la misma edad que él, así que saca los valores de los compañeros que tienen 17 años y tiene lo siguiente

Paga	5	6	8	9	10	12	15	20	25	30
Edad = 17	0	0	0	2	1	3	2	0	1	0

Como este chico tiene una paga de 10 euros, decide que la mayoría de sus compañeros tienen más paga que él, así que que intentará conseguir que su padre le suba la paga.

Lo que acabamos de calcular es la *distribución condicionada* de la variable paga fijado un valor de la edad, en este caso 17. De nuevo lo que hemos obtenido es una variable unidimensional a la que podemos calcularle las medidas de centralización y dispersión que ya conocemos.

Ejercicio 2.4.3 *Calcula la tabla de frecuencias de la variable edad para paga=15 euros.*

Ejercicio 2.4.4 *Calcula la tabla de frecuencias, con las frecuencias marginales, para el peso y la respuesta a la pregunta 3.1.*

2.5 Diagramas de dispersión o nubes de puntos

Como en el caso de las variables unidimensionales, en muchas ocasiones, los datos se interpretan de manera más sencilla si los representamos de forma gráfica. Como en el caso de las variables unidimensionales, en muchas ocasiones, los datos se interpretan de manera más sencilla si los representamos de forma gráfica. De cualquier manera, ahora estamos ante otra situación, ya que necesitamos representar dos variables con sus correspondientes frecuencias. Para ello, el gráfico que utilizaremos es la *nube de puntos* o *diagrama de dispersión*. Vamos a ver cómo se construye: representamos en el eje de abscisas la variable paga y en el eje de ordenadas la variable edad. A los puntos que representamos le damos mayor grosor según la frecuencia con la que aparecen o bien dibujamos tantos puntos como indica la frecuencia.

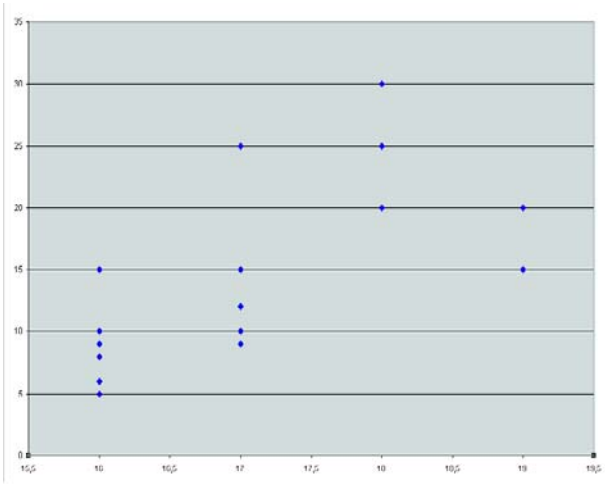

Figura 2.1: diagrama de dispersión

La forma de que tenga el diagrama de dispersión nos dará una idea de la posible dependenecia que haya entre las variables, como veremos a continuación.

Ejercicio 2.5.1 *Dibuja el diagrama de dispersión de la variable peso y las respuestas a la pregunta 3.1*

2.6 Dependencia funcional y dependencia estadística

Imagina que estudias los siguientes pares de variables:

- La altura de una persona y su número de pie
- La paga semanal y la altura
- El número de miembros de una familia y el número de habitaciones de su vivienda
- La altura desde la que se tira y el tiempo que tarda en caer un objeto deteminado
- El peso y el número de hermanos

Para cada una de estas situaciones nos interesa saber si existe o no relación entre las variables que medimos, si el valor de una influye sobre el de la otra. El caso 4, por ejemplo, es muy sencillo. Sabemos (por física) que hay una *relación funcional* entre ambas variables, una ecuación que las relaciona. En otros casos, podemos intuir que no hay ninguna relación, como en el caso 2 y el 5. Sin embargo, en los casos 1 y 3 existe la posibilidad de que exista una posible relación entre las variables que no somos a priori capaces de concretar.

Los diagramas de dispersión pueden tomar diferentes formas y pueden orientarnos mucho sobre cómo se comportan las variables. Los utilizaremos como primera orientación, aunque posteriormente veremos maneras más fiables de decidir cuándo dos variables están relacionadas.

Como ya hemos visto, hay distintos grados de relación entre variables. Decimos que existe *dependencia funcional* si nos encontramos en un caso similar al caso 4 que hemos visto anteriormente, es decir, Y depende funcionalmente de X cuando a cada valor x_i le podemos asignar un único valor y_j de manera que $y_j = f(x_i)$, esto es, cuando el valor de una variable determina exactamente el valor de la otra. La dependencia funcional será lineal cuando todos los pares de puntos se encuentren en una recta; será curvilínea cuando se encuentren en una curva definida por la función $y = f(x)$.

Dos variables X e Y se dicen *independientes* si el valor de una de ellas no influye sobre la otra, lo que significa que las distribuciones condicionadas relativas coinciden.

En el resto de los casos hablaremos de *dependencia o relación estadística*. Esta dependencia puede ser más o menos fuerte según los casos. Podemos tener una idea de si es fuerte o débil a través del diagrama de dispersión, observando que será más fuerte cuanto más se acerque la nube de puntos a la representación de una función.

Diagramas de dispersión en los que las variables tengan dependencia lineal o dispersión curvilínea pueden ser por ejemplo:

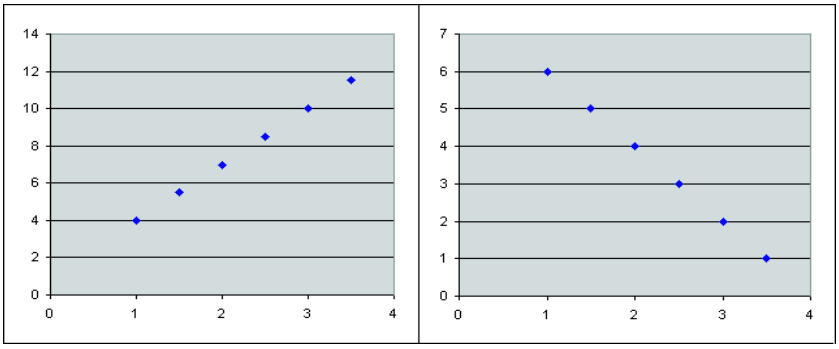

Figura 2.2: dependencia lineal

(Añadir diagramas de dispersión como ejemplos.)

Ejercicio 2.6.1 *¿Puedes deducir alguna conclusión sobre la posible dependencia entre las variables peso y la respuesta a la pregunta 3.1 a partir del diagrama de dispersión que dibujaste en la sección anterior?*

2.7 Covarianza

Recuerda el diagrama de dispersión de las dos variables que estamos estudiando. En principio, no resulta fácil deducir qué tipo de relación hay entre ellas, pero por ejemplo ¿crees que, en general, aumenta la paga al aumentar la edad? ¿o crees que es al revés? Intentamos ahora encontrar alguna

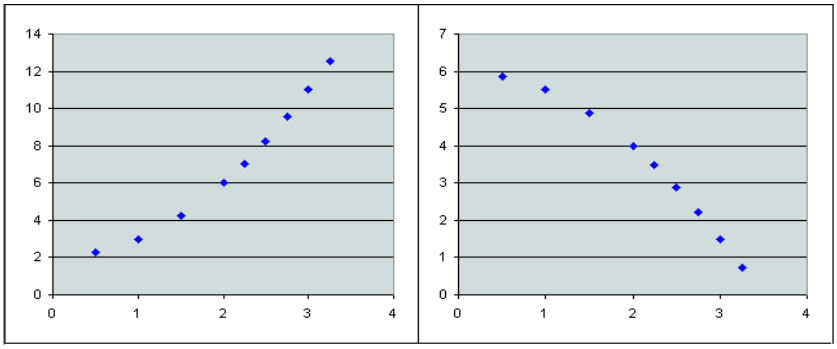

Figura 2.3: dependencia curvilínea

cantidad que nos dé una medida de si la relación entre dos variables es directa o inversa. Lo que utilizaremos será la *covarianza* que tiene la siguiente expresión:

$$S_{xy} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \overline{x})(y_j - \overline{y})n_{ij}}{n} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i\, y_j\, n_{ij}}{n} - \overline{x}\;\overline{y}$$

Esta cantidad también se conoce como varianza conjunta. Si la relación entre las variables es directa, la covarianza es positiva y si la covarianza es negativa la relación será inversa. Sabiendo que la media de edad es $16,8\widehat{6}$ y que la media de la paga es 13, para nuestro caso, $S_{xy} = 4,5\widehat{3}$ luego la relación es directa y la covarianza bastante alta.

Si te fijas en la expresión de la covarianza, su signo depende de las diferencias $(x_i - \overline{x})$ e $(y_j - \overline{y})$. Vamos a ver qué ocurre con la covarianza en algunos casos. Representamos 3 diagramas de dispersión, en los que marcamos el punto $(\overline{x}, \overline{y})$ que es el centro de gravedad de las distribuciones (ver figura 2.4):

Ocurre que en el gráfico 2 tendremos covarianza alta, puesto que las diferencias $(x_i - \overline{x})$ e $(y_j - \overline{y})$ son siempre del mismo signo (x_i e y_j están en el primer y tercer cuadrante definidos por los ejes centrados en $(\overline{x}, \overline{y})$. Al ser estas diferencias del mismo signo, contribuyen de forma positiva a la suma.

En los otros dos casos, el 1 y el 3, no existe relación lineal y habrá tanto sumandos positivos como negativos, ya que los puntos aparece en los cuatro cuadrantes, lo que hará que se anulen unos con otros y el resultado sea más próximo a 0.

Puedes observar que la covarianza es una medida que depende de las unidades, como en el caso unidimensional dependían la varianza y la desviación típica, por lo que debemos buscar otra medida que sea adimensional y nos permita comparaciones globales entre distribuciones.

Ejercicio 2.7.1 *Calcula la covarianza de las variables peso y respuesta a la pregunta 3.1. ¿qué podemos decir sobre la relación entre ellas a la vista de este valor?*

2.8 Correlación lineal

Buscamos ahora una medida que nos indique el grado de relación entre dos variables (de forma

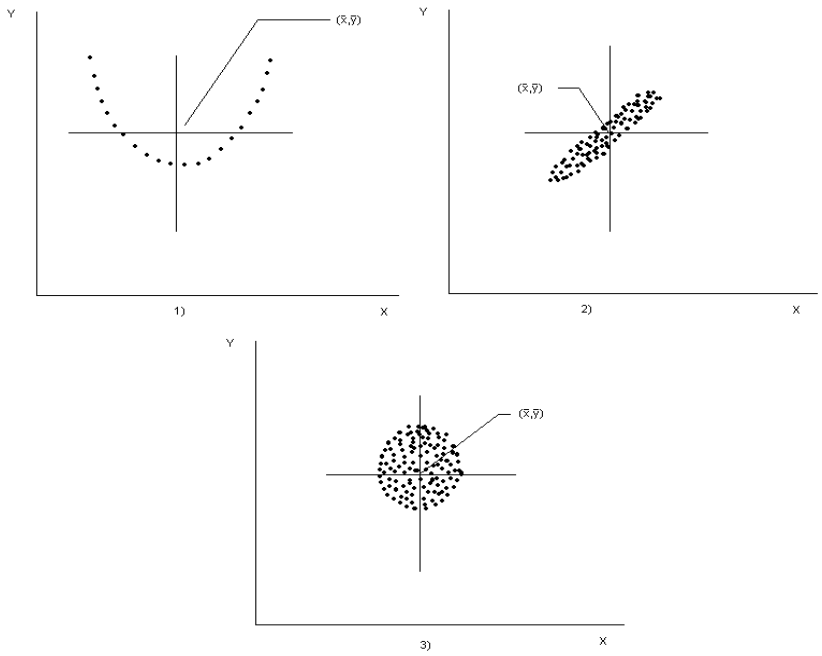

Figura 2.4: covarianzas

directa o inversa) y que no dependa de las unidades. Queremos además que nos mida el grado de relación lineal entre las dos variables.

Partimos de la covarianza que acabamos de ver, que si te fijas, depende del producto de las unidades en que están medidas las variables, ya que $(x_i - \overline{x})$ depende de las unidades de x_i e $(y_j - \overline{y})$ depende de las unidades de y_j, mientras que n_{ij} y n son adimensionales. Debemos dividir S_{xy} entre alguna cantidad que anule dichas unidades de medida. No conocemos ninguna otra medida de caracter bidimensional, así que pensemos en las medidas unidimensionales de cada una de las variables. Si recuerdas, la varianza de una variable depende de las unidades de dicha variable al cuadrado, luego no podemos utilizarla, pero la desviación típica de una variable depende de las unidades en las que están medidas estas variables. Esto quiere decir que el producto $S_x S_y$ depende del producto de las unidades de x por las unidades de y y es la cantidad que buecábamos como denominador. Entonces, definimos el *coeficiente de correlación lineal* de la siguiente manera

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

Vamos a calcularlo en nuestro caso, para las dos variables que tenemos. Sabemos que $S_{xy} = 4,5\widehat{3}$ y que $S_x = 1,008$ y que $S_y = 6,368$ luego $r = 0,706$, pero ¿qué significa este valor?

El valor de r está siempre entre -1 y 1. Si el valor de r está próximo a -1 o 1, entonces la dependencia lineal de las dos variables es fuerte, siendo directa si está próximo a 1 e inversa si está próximo a -1.

Si el valor de r está próximo a 0, la dependencia es débil en caso de que la haya. Si el valor coincide con -1 o 1, la dependencia es lineal y todos los puntos de la nube pertenecen a una recta.

Entonces, en nuestro caso, confirmamos que la relación es directa y como el valor de r es algo más de $0,7$ podemos decir que la dependencia lineal es considerable.

Ejercicio 2.8.1 *Calcula el coeficiente de correlación lineal de las variables peso y respuesta a la pregunta 3.1. ¿qué podemos decir sobre la relación entre ellas a la vista de este valor?*

2.9 Rectas de regresión

Imagina que sabes que un chico del instituto tiene una paga de 18 euros, pero no sabes su edad. Se nos podría ocurrir plantearnos la posibilidad de predecir el valor que puede tener la edad de este chico. ¿Cómo podríamos hacerlo? Hemos hablado durante todo el capítulo de la posible relación entre las dos variables, así que es el momento de utilizarla. Si fuéramos capaces de escribir la ecuación que relaciona la edad con la paga, sólo tendríamos que sustituir y obtendríamos el valor que buscamos.

Pero, desafortunadamente, no es tan sencillo. Como conocemos el hecho de que la correlación lineal entre las variables es razonablemente grande, podemos intentar encontrar la recta que mejor se ajuste a los puntos y luego sustituir el valor de la paga para obtener la edad. Esta recta es la que conocemos como *recta de regresión*. Vamos a ver cómo la definimos para posteriormente calcular la que corresponde a nuestro ejemplo.

Dadas dos variables X e Y, se define la recta de regresión como la recta que hace mínima la suma de los cuadrados de las distancias de los puntos observados a los puntos estimados.

Para la recta de regresión de Y sobre X, que será de la forma $y = ax + b$ se hace mínima la suma de los cuadrados de las distancias entre los puntos observados y_i y las ordenadas previstas por la recta para dichos puntos $ax_i + b$. La ecuación de esta recta viene dada por:

$$Y - \overline{y} = \frac{S_{xy}}{S_x^2}(X - \overline{x})$$

Utilizaremos esta recta cuando queramos estimar el valor de Y una vez conocido el valor de X

En el caso de la recta de regresión de X sobre Y que será de la forma $x = cy + d$ se hace mínima la suma de cuadrados de las distancias entre los puntos observados x_i y la predicción para las abscisas de esos puntos, $cy_i + d$. La ecuación de esta recta sería

$$X - \overline{x} = \frac{S_{xy}}{S_y^2}(Y - \overline{y})$$

Utilizaremos esta recta cuando queramos predecir el valor de X una vez conocido el de Y.

Calculemos ahora la recta de regresión para el caso práctico que estamos tratando. Como nuestras variables son la paga (X) y la edad (Y) lo que debemos calcular es la recta de regresión de Y sobre X. Tenemos que

$$\overline{x} = 13 \qquad \overline{y} = 16,8\widehat{6} \qquad S_{xy} = 4,5\widehat{3} \qquad S_x = 6,368 \qquad S_x^2 = 40,551$$

luego nuestra recta es

$$Y - 16,8\widehat{6} = \frac{4,5\widehat{3}}{40,551}(X - 13)$$

o lo que es lo mismo

$$Y - 16,8\widehat{6} = 0,111(X - 13) \Rightarrow Y = 0,111X + 15,413$$

luego si la paga de este chico es $x = 18$, su edad debe ser:

$$Y = 0,111 \cdot 18 + 15,413 = 17,42$$

es decir, este chico tiene 17 años.

Debemos hacer algunas puntualizaciones sobre la recta de regresión. Lo primero es que el punto de corte de las dos rectas de regresión (la de X sobre Y y la de Y sobre X) es $(\overline{x}, \overline{y})$, salvo en el caso de correlación lineal 1 o -1, caso en el que las dos rectas coinciden.

Si queremos realizar estimaciones con la recta de regresión, tenemos que tener considerar que se dan alguna de las siguientes circunstancias:

- Que observando el diagrama de dispersión podamos deducir una posible relación lineal entre las variables.
- Que el coeficiente de correlación lineal esté próximo a -1 o a 1.
- Que el sentido común nos indique que existe una posible relación lineal entre las variables.

Una manera alternativa de expresar la recta de regresión es la siguiente:

- Para el caso de la recta de regresión de Y sobre X, ésta es de la forma $y = ax + b$ donde

$$a = \frac{S_{xy}}{S_x^2} \qquad b = \overline{y} - \frac{S_{xy}}{S_x^2}\overline{x}$$

- Para el caso de la recta de regresión de X sobre Y, ésta es de la forma $x = cy + d$ donde

$$c = \frac{S_{xy}}{S_y^2} \qquad d = \overline{x} - \frac{S_{xy}}{S_y^2}\overline{y}$$

Ejercicio 2.9.1 *Calcula las rectas de regresión para las variables peso y la respuesta a la pregunta 3.1. Si una persona pesa 67 kg ¿puedes estimar cuál será su respuesta a la pregunta 3.1?*