8

Estadística

8.1. Distribuciones unidimensionales

Tablas de frecuencias

En este tema nos ocuparemos del tratamiento de datos estadísticos. Nuestro objeto de estudio será pues el valor de una cierta variable estadística en una cierta población (por ejemplo, la altura de los alumnos de una clase, o el número de automóviles en cada provincia española).

Los datos no siempre se toman sobre toda la población que se quiere estudiar - muchas veces es simplemente imposible- sino sobre una determinada muestra, que se considera representativa. Nosotros utilizaremos la letra N para denotar el número total de elementos de nuestra población o muestra.

Una forma de presentar los datos es a través de una tabla de frecuencias. En la columna de la izquierda se escribe cada uno de los posibles valores de la variable estadística puede tomar x_i. Junto a x_i se escribe su **frecuencia absoluta** n_i, esto es, las veces que se repite el dato x_i. A su vez, también podemos incluir una columna con las **frecuencias relativas**, f_i que se obtienen como resultado de dividir la frecuencia absoluta entre la población total $f_i = \frac{n_i}{N}$.

Ejemplo 8.1. Las edades de los niños de una unidad infantil son

$$1, 2, 2, 3, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 2, 1, 1, 4, 1$$

Puesto que tenemos 20 niños, el tamaño total de la población es $N = 20$. Hay 7 niños de 1 año, 6 niños de 2 años, 4 niños de 2 años y 3 niños de 4 años. La tabla de frecuencias es la siguiente:

x_i	n_i	f_i		
1	7	7/20	=	0,35
2	6	6/20	=	0,3
3	4	4/20	=	0,2
4	3	3/20	=	0,15
	20			1

Las frecuencias relativas a veces se expresan en porcentajes, que se obtienen multiplicando cada f_i por 100. En este caso serían 35 % de niños de 1 año, 30 % de 2 años, 20 % de 3 años y 15 % de 4 años.

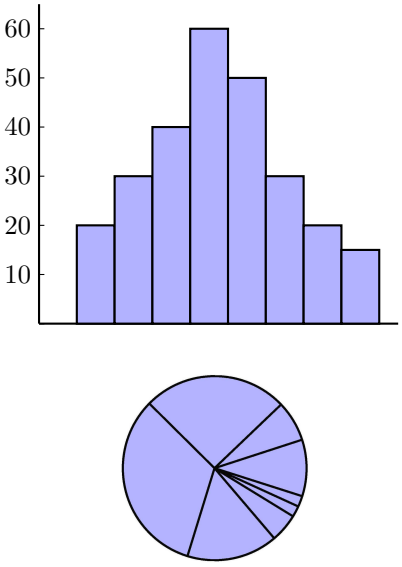
A veces podemos tener una variable continua, o que toma demasiados valores como para que el esquema anterior de frecuencias resulte operativo. En este caso, dividimos el rango de valores en clases.

Ejemplo 8.2. Imaginemos una clase con 88 estudiantes a los que se les pone un test con una puntuación de 0 a 100. No tendría mucho sentido contar cuántos alumnos han sacado cada nota en particular. En lugar de eso, si la menor nota ha sido 38 y la mayor nota 79, dividimos el intervalo $[38, 79)$ en siete subintervalos iguales que llamamos **clases**, y la frecuencia absoluta n_i será ahora cuántos alumnos han sacado una nota en cada una de esas clases. A cada clase se le asigna una **marca de clase**, que sería la nota media de cada intervalo. De esta forma, simplificamos los datos, asumiendo que todos los alumnos que han sacado entre 38 y 43 han sacado 41, los que han sado entre 44 y 49, suponemos que han sacado 47, etc.

Clases	Marca	n_i	f_i		
[38-44)	41	7	7/88	$\approx$	8.0 %
[44-50)	47	8	8/88	$\approx$	9.1 %
[50-56)	53	15	15/88	$\approx$	17.0 %
[56-62)	59	25	25/88	$\approx$	28.4 %
[62-68)	65	18	18/88	$\approx$	20.5 %
[68-74)	71	9	9/88	$\approx$	10.2 %
[74-80)	77	6	6/88	$\approx$	6.8 %
		88			100 %

Así pues, por ejemplo, leemos en la tabla que ha habido 15 alumnos con nota entre 50 y 55, lo que supone aproximadamente un 17 % del total de alumnos.

Las representaciones gráficas más habituales son de diagramas de barras (o histograma) y sectores. El área de cada barra o sector es proporcional a la correspondiente frecuencia.



Medidas de centralización: Media, mediana y moda

Una medida de centralización es un valor que se asocia a una distribución, que represente un promedio de todos los valores que toma la variable.

La **media aritmética** es la más utilizada: el cociente entre la suma de todos los datos y el número de ellos.

$$\overline{X} = \frac{x_1 + \cdots + x_N}{N}$$

Equivalentemente, si n_i es la frecuencia absoluta de cada valor x_i,

$$\overline{X} = \frac{n_1 x_1 + \cdots + n_k x_k}{N}$$

En las distribuciones de datos agrupados los valores x_i corresponden a las marcas de clase.

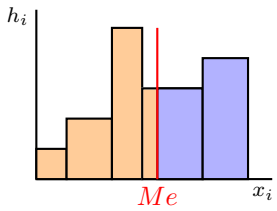
Ejemplo 8.3. La media del test realizado a 88 alumnos que consideramos en un ejemplo anterior sería 59.14:

Clases	Marca	n_i	$n_i x_i$
[38-44)	41	7	287
[44-50)	47	8	376
[50-56)	53	15	795
[56-62)	59	25	1 475
[62-68)	65	18	1 170
[68-74)	71	9	639
[74-80)	77	6	462
		88	5 204

$$\overline{X} = \frac{5\,204}{88} = 59.14$$

Se llama **mediana** de una distribución, y se designa por Me, al número tal que, ordenados los datos de forma creciente o decreciente, la mitad son inferiores a dicho número y la otra mitad son superiores. Si N es un número impar, existe un único valor de la variable en el centro de la distribución, y éste es la mediana. En el caso de que N sea par, la mediana se define como la media aritmética de los dos valores centrales.

Para distribuciones de datos agrupados, la **mediana** es el valor cuya vertical divide el histograma en partes de igual superficie,



Para calcular la mediana en estos casos, se suman las frecuencias absolutas hasta encontrar el valor que iguale o supere a $N/2$: este valor es el intervalo mediano. Si este intervalo es $[L_{i-1}, L_i)$, $a_i = L_i - L_{i-1}$ es su amplitud, y N_{i-1} es la suma de las frecuencias hasta el intervalo anterior, entonces la mediana viene dada por:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i$$

Se define la **moda** de una distribución estadística, y se designa por Mo, como el valor de la variable al que corresponde mayor frecuencia. Puede resultar que hay uno o más valores con la misma frecuencia máxima, por lo que se habla de distribuciones bimodales, trimodales, etc.

Para distribuciones de datos agrupados, el intervalo al que corresponde mayor altura en el histograma se llama intervalo modal. Puede tomarse como moda simplemente el punto medio del intervalo modal, aunque a veces se utilizan otros criterios.

Ejemplo 8.4. Supongamos que al preguntar sobre el número de hijos a un grupo de mujeres obtenemos los siguientes resultados:

$$1, 0, 0, 2, 3, 2, 1, 0, 2, 2, 0, 2, 8, 1, 1, 0, 2, 0, 1, 2, 2, 1, 0, 3, 2$$

La media de la distribución la obtenemos dividiendo el número total de hijos entre el de mujeres: $38/25 = 1.52$ hijos por mujer. Si ordenamos los datos de menor a mayor, observamos que la mediana es 1, ya que la posición central la ocupa el número 1, dejando doce valores a la izquierda y doce a la derecha.

$$0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, \mathbf{1}, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 8$$

Si añadimos un dato de una mujer con dos hijos, tendremos un total de 26 datos, número par. Entonces no habría un valor central, sino dos y la mediana sería la media de esos dos valores centrales: en ese caso 1.5.

$$0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, \mathbf{1,2}, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 8$$

En cualquiera de los casos, la moda de la distribución vale 2, puesto que es el valor que más se repite.

Medidas de dispersión: varianza y desviación típica

Las medidas de dispersión de una distribución estadística nos indican si los valores que toma están muy alejados o muy próximos entre sí. La medida de dispersión más sencilla es el **recorrido** o **rango**: la diferencia entre los valores máximo y mínimo de la variable.

La **varianza** de una distribución es la media aritmética de los cuadrados de las desviaciones de los datos respecto a su media aritmética.

$$S^2 = \frac{\sum_{n=1}^{k}(x_i - \overline{X})^2 n_i}{N} = \frac{\sum_{n=1}^{k} x_i^2 n_i}{N} - \overline{X}^2$$

Se llama **desviación típica** a la raíz cuadrada, con signo positivo, de la varianza; se representa por S y es

$$S = +\sqrt{S^2}$$

La desviación típica representa cuánto suelen alejarse los datos de la distribución de su media aritmética, como promedio. Una desviación típica baja indica que en general los datos están muy cercanos de la media aritmética, mientras que un valor alto indica una predominancia de datos que toman valores alejados de la media.

Ejemplo 8.5. Consideremos de nuevo el caso del notas del test

Clases	Marca	n_i	$n_i x_i$
[38-44)	41	7	287
[44-50)	47	8	376
[50-56)	53	15	795
[56-62)	59	25	1 475
[62-68)	65	18	1 170
[68-74)	71	9	639
[74-80)	77	6	462
		88	5 204

Habíamos calculado la media $\overline{X} = \frac{5\,204}{88} = 59.14$. En la siguiente tabla, escribimos las desviaciones de cada valor respecto de la media (es decir, la diferencia entre cada valor de marca y la media) y sus cuadrados:

Clases	x_i	n_i	$x_i - \overline{X}$	$(x_i - \overline{X})^2$
[38-44)	41	7	-18.14	329.06
[44-50)	47	8	-12.14	147.38
[50-56)	53	15	-6.14	37.70
[56-62)	59	25	-0.14	0.02
[62-68)	65	18	5.86	34.34
[68-74)	71	9	11.86	140.66
[74-80)	77	6	17.86	318.98

La varianza es el promedio de de las desviaciones al cuadrado, es decir

$$S^2 = \frac{7 \cdot 329.06 + 8 \cdot 147.38 + \cdots}{88} \approx 89.16$$

y la desviación típica, la raíz cuadrada de la varianza

$$S = \sqrt{89.16} \approx 9.44$$

8.2. Distribuciones bidimensionales

En una distribución bidimensional se consideran dos variables estadísticas sobre una misma población. Las representaremos, en general, por (X, Y).

Estas distribuciones suelen presentarse mediante una tabla de tres columnas, apareciendo en las dos primeras los valores de las variables, y en la tercera, la frecuencia del par correspondiente, es decir, en la forma

X	Y	n_{ij}
x_1	y_1	n_{11}
x_1	y_2	n_{12}
$\vdots$	$\vdots$	$\vdots$
x_i	y_j	n_{ij}
$\vdots$	$\vdots$	$\vdots$
x_h	y_k	n_{hk}

En ocasiones es preferible hacerlo mediante una *tabla de doble entrada*, con disposición rectangular, en la forma

$Y \setminus X$	x_1	x_2	x_3	$\cdots$	x_i	$\cdots$	x_h
y_1	n_{11}	n_{21}	n_{31}	$\cdots$	n_{i1}	$\cdots$	n_{h1}
y_2	n_{12}	n_{22}	n_{32}	$\cdots$	n_{i2}	$\cdots$	n_{h2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
y_j	n_{1j}	n_{2j}	n_{3j}	$\cdots$	n_{ij}	$\cdots$	n_{hj}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
y_k	n_{1k}	n_{2k}	n_{3k}	$\cdots$	n_{ik}	$\cdots$	n_{hk}

Ejemplo 8.6. Supongamos que contamos el número de pizzerías (X) y el número de hamburgueserías (Y) en 80 localidades de una región. Obtenemos la siguiente tabla de frecuencias:

X	Y	n_{ij}
0	1	4
1	1	3
1	3	4
2	0	2
2	2	9
2	3	3
3	1	6
3	2	12
3	3	5
3	4	2
4	0	2
4	1	7
4	2	15
4	4	1
5	2	5

La interpretación es que hay 4 localidades que tienen 0 pizzerías y 1 hamburguesería, hay 12 localidades con 3 pizzerías y 2 hamburgueserías, etc. La tabla de doble entrada sería la siguiente:

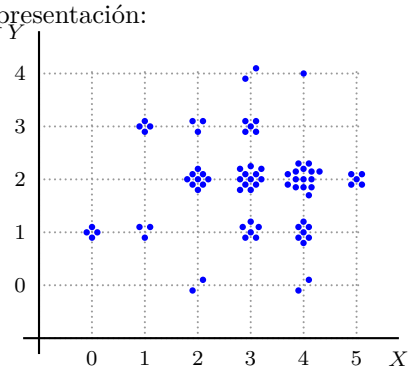
Y \ X	**0**	**1**	**2**	**3**	**4**	**5**
0	0	0	2	0	2	0
1	4	3	0	6	7	0
2	0	0	9	12	15	5
3	0	4	3	5	0	0
4	0	0	0	2	1	0

La forma más usual de representar gráficamente las distribuciones bidimensionales es el **diagrama de dispersión** o **nube de puntos**, que se obtiene al considerar dos ejes coordenados, situando en el eje horizontal los valores de la variable X y en el vertical los de la variable Y; en las proximidades del par (x_i, y_j) se colocan tantos puntos como indica su frecuencia conjunta n_{ij}.

Ejemplo 8.7. La distribución bidimensional que consideramos anteriormente

Y \ X	**0**	**1**	**2**	**3**	**4**	**5**
0	0	0	2	0	2	0
1	4	3	0	6	7	0
2	0	0	9	12	15	5
3	0	4	3	5	0	0
4	0	0	0	2	1	0

tendría la siguiente representación:



También se puede optar por utilizar puntos de distinto tamaño según la frecuencia, o simplemente representar un punto por cada valor (x_i, y_i), en los casos en que las variables varían continuamente y no hay repeticiones.

Covarianza

Para una distribución estadística bidimensional (X, Y), se llama **covarianza** a la media aritmética de los productos de las desviaciones de cada variable respecto a su media aritmética; se indicará por S_{XY}, y está dada por la fórmula

$$S_{XY} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \overline{X})(y_j - \overline{Y})n_{ij}}{N}$$

Se puede calcular más fácilmente en la forma:

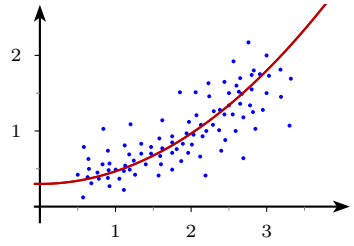
$$S_{XY} = \frac{\sum_i \sum_j x_i y_j n_{ij}}{N} - \left(\frac{\sum_i \sum_j x_i n_{ij}}{N}\right)\left(\frac{\sum_i \sum_j y_j n_{ij}}{N}\right)$$

es decir, la covarianza es igual a *la media de los productos menos el producto de las medias.*

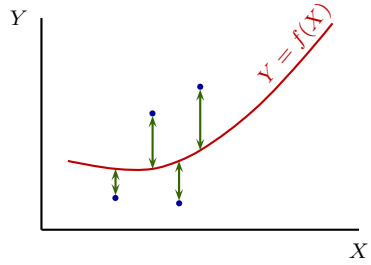
Regresión

Dada una distribución bidimensional, cabe preguntarse si las dos variables son independientes o si están relacionadas entre sí. Y si están relacionadas, cuál es esa relación. Por ejemplo, si consideramos sobre un grupo de personas la estatura (X) y el sueldo mensual (Y), lo lógico es pensar que se trata de dos variables completamente independientes. Sin embargo, si consideramos la estatura (X) y el peso (Z) sí que va que va a haber una relación importante, ya que las personas más altas suelen por lo general tener mayor peso.

La relación entre dos variables puede observarse al representar gráficamente la nube de puntos. Cuando dos variables están relacionadas, la nube de puntos tiende a concentrarse en torno a la gráfica de una determinada función. El problema de la regresión consiste en encontrar esa función.



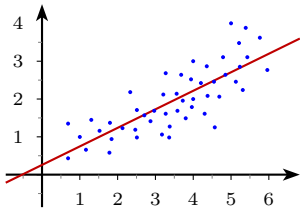
El método más habitual es la **regresión por mínimos cuadrados**. Primero hemos de decidir qué tipo de función creemos que es la más apropiada para nuestro caso. Por ejemplo, podemos decidir aproximar por un polinomio de segundo grado, $f(x) = ax^2 + bx + c$. El siguiente paso consistiría en encontrar los valores de a, b y c que hacen que la diferencia entre la gráfica de la función y la distribución sea lo más pequeña posible. Esta diferencia se cuantifica calculando, para cada valor de la distribución (x_i, y_i) la diferencia $y_i - f(x_i)$, y después sumando los cuadrados de todas esas diferencias: $\sum_{i=1}^{N}(y_i - f(x_i))^2$.



En la figura, los puntos azules son los de la gráfica de la distribución. El proceso de regresión por mínimos cuadrados consiste en encontrar la función f cuya gráfica tiene la propiedad de que la suma de los cuadrados de las longitudes indicadas en verde sea lo menor posible.

El caso más sencillo es la **regresión lineal**, cuando el tipo de función por el que aproximamos es una función lineal $f(x) = ax + b$. Es decir, se trata de encontrar una recta $y = ax + b$ que sea la que mejor aproxime a nuestra distribución bidimensional (X, Y) por mínimos cuadrados. Esta recta se halla mediante la siguiente fórmula:

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



donde recordamos que $\overline{X}$ y $\overline{Y}$ son las medias aritméticas de X e Y respectivamente, S_X es la varianza de X, S_{XY} es la covarianza de X e Y. La ecuación anterior nos indica que la recta hallada pasa por el punto $(\overline{X}, \overline{Y})$, llamado **centro de gravedad** de la distribución bidimensional, y tiene por pendiente $a = \frac{S_{XY}}{S_X^2}$, llamado el **coeficiente de regresión**. Esta recta se llama **recta de regresión de Y sobre X**. Es importante especificar el orden de las variables, puesto que la recta de regresión de Y sobre X no

coincide con la de X sobre Y. Ambas rectas aspiran a ser las que mejor aproximan la distribución, pero en un caso hemos tratado de minimizar distancias medidas en vertical y en otro caso en horizontal, por lo que el resultado no será exactamente el mismo.

Se llama **coeficiente de correlación lineal de Pearson** al valor

$$r = \frac{S_{XY}}{S_X \cdot S_Y}$$

Este coeficiente sirve para medir hasta qué punto la recta de regresión es una buena aproximación de la distribución: mejor cuanto más próximo esté el valor de $|r|$ a 1 y peor cuanto más se acerque a 0. Si $r = \pm 1$, la correlación lineal es perfecta, directa o inversa, es decir, la nube de puntos está situada, toda ella, sobre la recta de regresión, con pendiente positiva para $r = 1$ y negativa para $r = -1$. Si $r = 0$, no existe dependencia lineal entre las variables, pudiendo darse una dependencia no lineal, o bien puede ocurrir que las variables sean independientes.