Distribuciones unidimensionales y bidimensionales

Ordenación de datos

☞ Teniendo en cuenta que la estadística es la ciencia que trata la recopilación, organización, presentación, análisis e interpretación de datos, y en ocasiones su gran finalidad es la toma de decisión más adecuadas, es importante la recopilación, ordenación y tratamiento de datos obtenidos, que además es fundamental para poder extraer conclusiones.

Conviene también recordar que si $(x_1, x_2, \ldots, x_k)$ es una **m.a.s.** de una **variable estadística unidimensional X**, donde x_1 se repite n_1 veces, x_2 se repite n_2 veces, . . . , x_k se se repite n_k veces, con . Para cada $i = 1, 2, 3, \ldots, k$ denominamos

- n = frecuencia total de datos.
- n_i = frecuencia absoluta de cada elemento, individuo, modalidad o clase x_i .
- $f_i = \dfrac{n_i}{n}$ = frecuencia relativa de cada elemento, individuo, modalidad o clase x_i .
- $N_i = \sum_{j=1}^{i} n_i$ = frecuencia absoluta acumulada de cada individuo, modalidad o clase x_i .
- $F_i = \sum_{j=1}^{i} f_i$ = frecuencia relativa acumulada de cada individuo, modalidad o clase x_i .

Y la matriz de datos será de la forma

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
…	…	…	…	...
x_{k-1}	n_{k-1}	f_{k-1}	N_{k-1}	F_{k-1}
x_k	n_k	f_k	N_k	F_k

Ejemplo: Si se miden la altura en cm. de 44 personas de entre 20 y 30 años, redondeada un decimal, y una vez agrupados los datos en intervalos se obtienen los siguientes datos

x_i	n_i
148,9 – 157,8	*3*
157,8 -166,7	*9*
166,7 -175,6	*11*
175,6 -184,5	*9*
184,5 – 193,4	*7*
193,4 – 202,3	*5*

Teniendo en cuenta que en el caso de variables aleatorias continuas agrupados por intervalos tomamos como valor x_i *(marca de clase) el valor medio del intervalo* I_i *, podemos construir la matriz de datos siguiente:*

I_i	n_i	f_i	N_i	F_i
153,35	*3*	$\frac{3}{44}$	*3*	$\frac{3}{44}$
162,25	*9*	$\frac{9}{44}$	*12*	$\frac{12}{44}$
171,15	*11*	$\frac{11}{44}$	*23*	$\frac{23}{44}$
180,05	*9*	$\frac{9}{44}$	*32*	$\frac{32}{44}$
188,95	*7*	$\frac{7}{44}$	*39*	$\frac{39}{44}$
197,85	*5*	$\frac{5}{44}$	*44*	1

En el caso de variables estadísticas bidimensionales (X, Y) , como las muestras serán de la forma $((x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k))$, donde (x_1, y_1) se repite n_1 veces, (x_2, y_2) se repite n_2 veces, …. , (x_k, y_k) se repite n_k veces, con $\sum_{i=1}^{k} n_i = n$. Para cada $i \in \{1, 2, 3, \ldots, \}$ y en este caso n_i, f_i, N_i y F_i serán los datos de (x_i, y_i) . Y la matriz de datos será de la forma

x_i	y_i	n_i	f_i	N_i	F_i
x_1	y_1	n_1	f_1	N_1	F_1
x_2	y_2	n_2	f_2	N_2	F_2
…	...	…	…	…	...
x_{k-1}	y_{k-1}	n_{k-1}	f_{k-1}	N_{k-}	F_{k-1}
x_k	y_k	n_k	f_k	N_k	F_k

Ejemplo: Las calificaciones de 40 alumnos en Matemáticas y Física han sido las siguiente:

X = calificación Matemáticas	3	4	5	6	6	7	7	8	10
Y = calificación Física	2	5	5	6	7	6	7	9	10
Número de alumnos	4	5	12	4	5	4	2	1	2

Y podemos construir la matriz de datos siguiente:

x_i	y_i	n_i	f_i	N_i	F_i
3	*2*	*4*	$\frac{4}{40}$	*4*	$\frac{4}{40}$
4	*5*	*6*	$\frac{6}{40}$	*10*	$\frac{10}{40}$
5	*5*	*12*	$\frac{12}{40}$	*22*	$\frac{22}{40}$
6	*6*	*4*	$\frac{4}{40}$	*26*	$\frac{26}{40}$
6	*7*	*5*	$\frac{5}{40}$	*31*	$\frac{31}{40}$
7	*6*	*4*	$\frac{4}{40}$	*35*	$\frac{35}{40}$
7	*7*	*2*	$\frac{2}{40}$	*37*	$\frac{37}{40}$
8	*9*	*1*	$\frac{1}{40}$	*38*	$\frac{38}{40}$
10	*10*	*2*	$\frac{2}{40}$	*40*	1
Sumas		*40*	1		

Parámetros unidimensionales

☞ **Parámetros de centralización**.- Son los valores que nos aporta información de concentración de datos, siendo algunos de ellos

Media muestral: $\bar{x}=\sum_{i=1}^{k} f_i . x_i=\sum_{i=1}^{k} \frac{n_i . x_i}{n}$

Moda: $M_d=x_d \in \{x_1, x_2, \ldots, x_k\}$ tal que $Max_{i\in\{1,2,\ldots,k\}} f_i=f_d$

Es el valor o valores que tienen mayor frecuencia (*en caso de datos agrupados en intervalos se utilizan fórmulas de interpolación lineal*).

Percentiles: $p_{\frac{r}{c}}=x_t \in \{x_1, x_2, \ldots, x_k\}$ tal que $\sum_{i=1}^{t-1} f_i \le \frac{r}{c} \le \sum_{i=1}^{t} f_i$

Si $\frac{r}{c}=\frac{1}{2}$ se denomina Mediana M_e .

Si $c=4$, $p_{\frac{1}{4}}$, $p_{\frac{2}{4}}$ *y* $p_{\frac{3}{4}}$ *, se denominan cuartiles* Q_1 , Q_2 *y* Q_3 *respectivamente.*

Si $c=10$ *, los percentiles se denominan deciles y c = 100 centiles.*

La Mediana M_d (= Q_3) es el primer valor de la variable tal

que el número de valores menores que él es igual al número de valores mayores que él (*en caso de datos agrupados en intervalos se utilizan fórmulas de interpolación lineal*).

☞ **Parámetros de dispersión**.- Son los valores que nos proporcionan información de como están concentrados los datos de la distribución en torno a algún parámetro habitualmente central, siendo algunos de ellos

Rango o Recorrido: $R = x_{(n)} - x_{(1)}$ donde $x_{(n)} = max_{i \in \{1,2,\dots,k\}}$ y $x_{(1)} = min_{i \in \{1,2,\dots,k\}}$

El rango es la diferencia entre el mayor y el menor valor de los datos

Varianza s^2: $$s^2 = \sum_{i=1}^{k} f_i . (x_i - \bar{x})^2 = \sum_{i=1}^{k} \frac{n_i}{n} . (x_i - \bar{x})^2 = \sum_{i=1}^{k} \frac{n_i . x_i^2}{n} - (\bar{x})^2$$

Desviación típica s: $s = \sqrt{s^2}$

Coeficiente de Variación de Pearson: $C.V. = \frac{s}{|\bar{x}|}$

Este coeficiente es útil para comparar diversas distribuciones de datos.

Ejemplo.- Tras encuestar a 25 familias sobre el número de hijos que tenían se obtuvo los siguientes resultados

Número de hijos	x_i	0	1	2	3	4
Frecuencia absoluta	n_i	5	6	8	4	2

Pudiendo obtener la siguiente tabla de frecuencias, ampliando las columnas de datos $x_i . n_i$ ó $x_i . f_i$ y $x_i^2 . f_i$ ó $x_i^2 . f_i$, para poder calcular más cómodamente los parámetros

x_i	n_i	f_i	N_i	F_i	$x_i . n_i$	$\frac{x_i . n_i}{n}$	$x_i^2 . n_i$	$\frac{x_i^2 . n_i}{n}$
0	**5**	**0,20**	**5**	**0,2**	**0**	**0**	**0**	**0**
1	**6**	**0,24**	**11**	**0,44**	**6**	**0,24**	**6**	**0,24**
2	**8**	**0,32**	**19**	**0,64**	**16**	**0,64**	**32**	**1,28**
3	**4**	**0,16**	**23**	**0,48**	**12**	**0,48**	**36**	**1,44**
4	**2**	**0,08**	**25**	**0,32**	**8**	**0,32**	**32**	**1,28**
SUMAS	**25**	**1**			**34**	**1,68**	**106**	**4,24**

Y podemos obtener algunos parámetros como:

Media aritmética $$\bar{x} = \frac{0 \times 5 + 1 \times 6 + 2 \times 8 + 3 \times 4 + 4 \times 2}{25} = 1{,}68$$

Moda $M_d = 2$

Mediana $M_e = 2 = Q_2$

Recorrido $R=4-0=4$

Varianza $s^2=\frac{(0-1{,}68)^2+(1-1{,}68)^2+(2-1{,}68)^2 6+(3-1{,}68)^2+(4-1{,}68)^2}{25}=1{,}1418$

Desviación típica $s=\sqrt{s^2}=1{,}19$

Parámetros bidimensionales

En el caso de variables estadísticas bidimensionales (X,Y), además de poder hallar los parámetros unidimensionales para cada variable, como por ejemplo

	Variable X	Variable Y
Media	$\bar{x}=\frac{\sum_{i=1}^{k} x_i . n_i}{n}$	$\bar{y}=\frac{\sum_{i=1}^{k} y_i . n_i}{n}$
Varianza	$s_x^2=\frac{\sum_{i=1}^{k} (x_i-\bar{x})^{2.} n_i}{n}$	$s_y^2=\frac{\sum_{i=1}^{k} (y_i-\bar{x})^{2.} n_i}{n}$
Desviación típica	$s_x=\sqrt{s_x^2}$	$s_y=\sqrt{s_y^2}$

Podemos hallar parámetros bidimensionales como

- **La covarianza o varianza conjunta de** (X,Y) es

$$S_{xy}=\sum_{i=1}^{k}\frac{(x_i-\bar{x}).(y_i-\bar{y}).n_i}{n}=\sum_{i=1}^{k}\frac{x_i.y_i.n_i}{n}-\bar{x}.\bar{y}$$

- **El coeficiente de correlación** (X,Y) **es** $r=\frac{S_{xy}}{S_x.s_y}$

Concepto general de correlación

Se denomina correlación a la teoría que trata de estudiar la relación o dependencia entre varias variables estadísticas o de probabilidad, que en el caso particular de dos variables (X,Y), decimos:

- Existe una **correlación lineal entre X e Y**, si los valores (x_i , y_i) están próximos a una recta o curva determinada. Además, decimos que es **positiva o directa**, cuando la recta tiene pendiente positiva, y es **negativa o inversa**, cuando tiene pendiente negativa.
- Si no existe ninguna relación entre las variables X e Y, la **correlación es nula.**
- La **correlación de tipo funcional**, cuando existe una función que satisface todos los valores de la distribución.

Interpretación del coeficiente de correlación lineal.

Teniendo en cuenta que el signo del coeficiente de correlación bien determinado por el signo de la covarianza (*ya que las desviaciones típicas son siempre positivas*). Así pues, si la covarianza es positiva, la correlación es directa, si es negativa, es inversa, y si es nula, no existe correlación.

Además, teniendo en cuenta que se cumple

$$-1 \le r = \frac{S_{xy}}{S_x . s_y} \le 1$$

el tipo de dependencias que se crea entre las variables X e Y, según el valor r es

- Si $|r| = 1$ todos los valores de la variable bidimensional (X,Y) se encuentran situados sobre una recta; en consecuencia, satisface la ecuación de una recta. Entonces, se dice que entre las variables X e Y existe una dependencia funcional.
- Si $0 < |r| r < 1$, decimos que existe una dependencia lineal aleatoria, que será más fuerte cuando más próximo este r a -1 ó a 1 , y será más débil, cuando más próximo esté r a cero.
- Si r = 0, las variables X e Y son aleatoriamente independientes y no existe ningún tipo de relación funcional.

Estudio analítico de la regresión lineal

Si entre las variables estadísticas X e Y existe una fuerte correlación, el diagrama de puntos se condensa en torno a una recta.

Si tomamos como variable independiente X e Y como dependiente, entonces, el problema consiste en encontrar la ecuación de la recta que mejor se ajuste a la nube de puntos $\{(x_i, y_i) : i = 1, 2, \dots, k\}$.

Uno de los métodos más utilizados para hallar la recta de regresión, es el método de los mínimos cuadrados, que consiste en hacer mínima la suma de los cuadrados de las diferencias entre los valores observados experimentalmente y los teóricos que se obtengan mediante la recta. De la aplicación de dicho método se deduce que la recta regresión pasa por el punto $(\bar{x}, \bar{y})$, y que su pendiente, denominada **coeficiente de regresión** viene dada por $\frac{S_{xy}}{S_x^2}$, luego la **recta de regresión lineal de y sobre x**, será

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

Que si sustituimos en esta ecuación los valores x, podemos obtener de forma aproximada, los valores esperados para la variable y, que denominamos estimaciones o previsiones.

Si tomamos como variable independiente Y y X como dependiente, entonces, entonces, la **recta de regresión lineal de x sobre y**, será

$$x-\bar{x}=\frac{s_{xy}}{s_y^2}(y-\bar{y})$$

Que si sustituimos en esta ecuación los valores y, podemos obtener de forma aproximada, los valores esperados para la variable x.

Ejemplo: Una compañía discográfica ha recopilado la siguiente información sobre el número de conciertos dados durante el verano por 15 grupos musicales y las ventas de discos de estos grupos (expresados en miles de disco), obteniendo la siguiente tabla:

Discos / conciertos	[10,30)	[30,40)	[40,80)
[1,5)	3	0	0
[5,10)	1	4	1
[10,20)	0	1	5

a) Calcular el número medio de discos vendidos por estos grupos.

b) ¿Cómo es el grado de dependencia del número de conciertos dados por el grupo y el número de discos vendidos?

c) Obtener la recta de regresión que explica la dependencia anterior

d) Si un grupo musical ha vendido 18 000 discos, ¿qué número de conciertos es previsible que dé?

Solución:

Transformamos la tabla doble entrada en una tabla simple de frecuencias, tomando las marcas de clase, y ampliando algunas columnas de datos que nos ayudarán a calcular las medias, desviaciones típicas, covarianza y el coeficiente de correlación

x_i	y_i	n_i	$x_i.n_i$	$y_i.n_i$	$x_i^2.n_i$	$y_i^2.n_i$	$x_i.y_i.n_i$
3	**20**	**3**	**9**	**60**	**27**	**1200**	**180**
7,5	**20**	**1**	**7,5**	**20**	**56,25**	**400**	**150**
7,5	**35**	**4**	**30**	**140**	**225**	**4900**	**1050**
7,5	**60**	**1**	**7,5**	**60**	**56,25**	**3600**	**450**
15	**35**	**1**	**15**	**35**	**225**	**1225**	**525**
15	**60**	**5**	**75**	**300**	**1125**	**18000**	**4500**
		15	**144**	**615**	**1714,5**	**29325**	**6855**

Obteniendo

$$\bar{x}=\frac{144}{15}=9{,}6 \qquad \bar{y}=\frac{615}{15}=41$$

$$s_x=\sqrt{\frac{1714{,}5}{15}-9{,}6^2}=4{,}7 \qquad s_y=\sqrt{\frac{29325}{15}-615^2}=16{,}55$$

$$s_{xy} = \frac{6855}{15} - 9{,}6\,.\,41 = 63{,}4 \qquad r = \frac{63{,}4}{4{,}7\,.\,16{,}55} = 0{,}814$$

Luego

a) El número medio de discos vendidos es 9600

b) Para conocer el grado de dependencia que existe entre las dos variables, calculamos el coeficiente de correlación lineal:

$r = 0{,}814$

Luego, la correlación es positiva y alta.

c) La ecuación de la recta de regresión de y sobre x es

$$y - 41 = \frac{63{,}4}{4{,}7^2}\,.\,(x - 4{,}6)$$

Es decir

$$y = 2{,}87x + 13{,}44$$

d) Si un grupo musical ha vendido 18 000 discos, para saber el número de conciertos, que estimamos que dará sustituimos en la ecuación de la recta $x = 18$ *y obtenemos*

$$y = 2{,}87\,.\,18 + 13{,}44 = 65{,}1 \Rightarrow y = 61 \text{ concierto}$$