Estadística

Variables estadísticas

Una variable estadística representa una cualidad, que puede tomar diferentes valores para cada individuo de la población a estudiar. Por ejemplo, la altura de una persona, el peso, las cualificaciones en un examen,…

El conjunto de valores obtenidos al estudiar una población, se denominan distribución de la variable estadística. Es habitual representar la distribución en una tabla de frecuencias. Por ejemplo, al estudiar las cualificaciones en una determinada asignatura, nos podemos encontrar:

x_i	f_i
0	0
1	2
2	3
3	2
4	4
5	5
6	3
7	2
8	2
9	1
10	1
	25

En la primera columna observamos los valores de la variable (cualificaciones posibles), mientras que en la segunda aparece el número de alumnos que ha obtenido dicha cualificación.

Nótese que en total hay 25 alumnos, lo cual se indica N=25 o $\sum f_i = 25$ (el símbolo Σ se lee sumatorio y sirve para indicar la suma de todos los valores)

Parámetros estadísticos

Media

$$\overline{X} = \frac{\sum x_i \cdot f_i}{n}$$

Varianza

$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 \cdot f_i}{n}$ también se puede calcular como $\sigma^2 = \frac{\sum x_i^2 \cdot f_i}{n} - \bar{x}^2$

Desviación típica

Es la raíz cuadrada de la varianza $\sigma = \sqrt{\sigma^2}$

Estos parámetros suelen calcularse a partir de la tabla de frecuencias:

x_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
0	0	0	0
1	2	2	2
2	3	6	12
3	2	6	18
4	4	16	64
5	5	25	125
6	3	18	108
7	2	14	98
8	2	16	128
9	1	9	81
10	1	10	100
	25	122	736

$$\overline{X} = \frac{122}{25} = 4{,}88$$

$$\sigma^2 = \frac{736}{25} - 4{,}88^2 = 5{,}6256$$

$$\sigma = \sqrt{5{,}625} \approx 2{,}37$$

Distribuciones bidimensionales

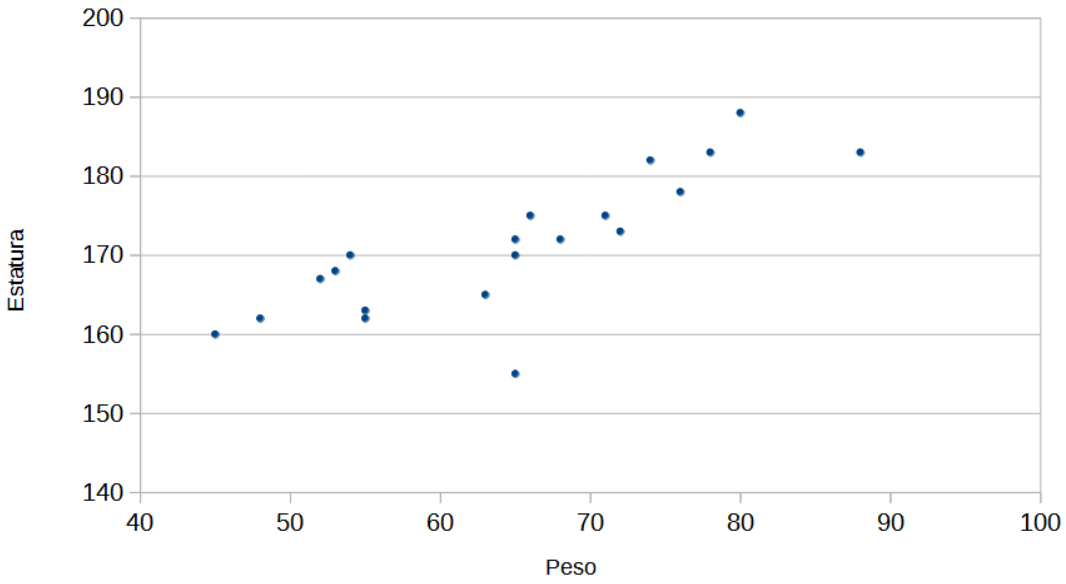
Si en una población estudiamos simultáneamente los valores de dos variables estadísticas, entonces hablaremos de una distribución bidimensional.

Ejemplo:

En un grupo de 20 personas, obtenemos estos valores para el peso y la estatura

Kg	65	80	88	52	55	71	68	45	55	72	66	54	65	78	63	48	53	76	65	74
Cm	155	188	183	167	163	175	172	160	162	173	175	170	172	183	165	162	168	178	170	182

Estos valores se pueden representar en lo que llamamos una **nube de puntos**.



Distribuciones marginales

Las distribuciones marginales se obtienen observando los datos de cada variable estadística por separado.

En el ejemplo anterior, llamando x al peso e y a la altura, tenemos:

La distribución de x

Kg	65	80	88	52	55	71	68	45	55	72	66	54	65	78	63	48	53	76	65	74

Con la cual podemos calcular su media y su varianza: $\overline{X}=64{,}65$ y $\sigma_x^2=127{,}23$

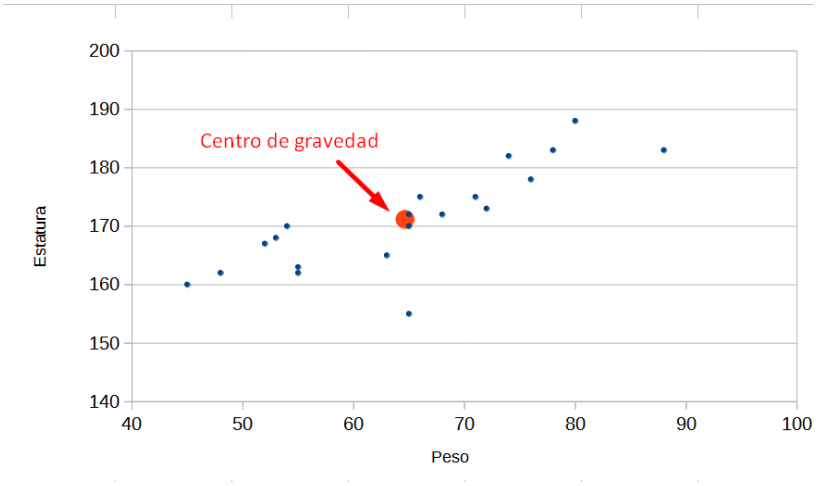
Y la distribución de y

Cm	155	188	183	167	163	175	172	160	162	173	175	170	172	183	165	162	168	178	170	182

Donde también calculamos media y varianza: $\overline{Y}=171{,}15$ y $\sigma_y^2=72{,}53$

Centro de gravedad

Es el punto cuyas coordenadas son $(\overline{X},\overline{Y})$



Covarianza

Se define la covarianza como el parámetro:

$$\sigma_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\sum x_i y_i}{N} - \bar{x} \cdot \bar{y}$$

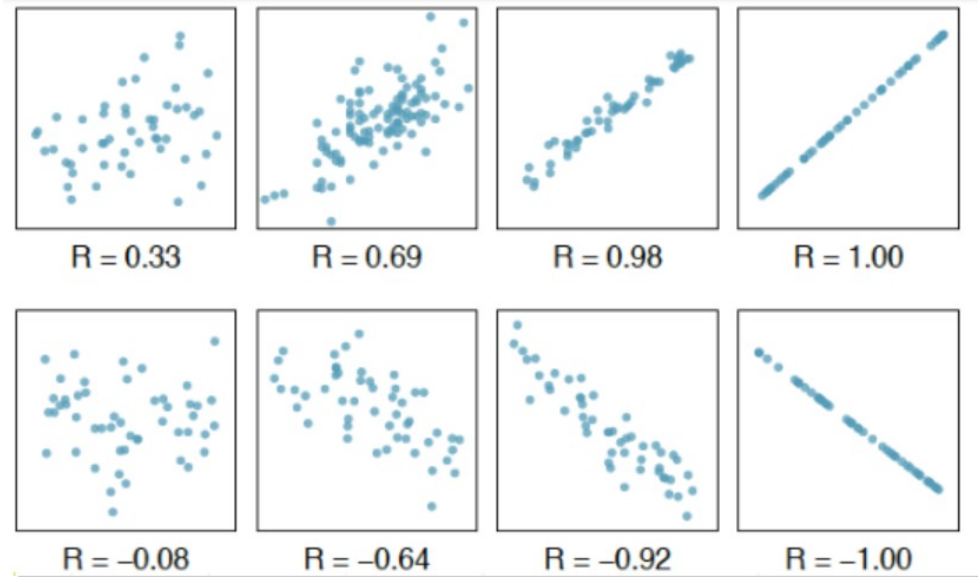
Coeficiente de correlación

El coeficiente de correlación se define como:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Propiedades

- $0 \leq |r| \leq 1$
- Si $|r| = 1$ entonces la alineación es perfecta (todos los puntos alineados). Si r=0 la alineación es nula.
- Si r>0 la alineación tiene pendiente positiva, mientras si r<0 la pendiente es negativa.



Ejemplo de cálculo:

20 datos

	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
	65	155	4225	24025	10075
	80	188	6400	35344	15040
	88	183	7744	33489	16104
	52	167	2704	27889	8684
	55	163	3025	26569	8965
	71	175	5041	30625	12425
	68	172	4624	29584	11696
	45	160	2025	25600	7200
	55	162	3025	26244	8910
	72	173	5184	29929	12456
	66	175	4356	30625	11550
	54	170	2916	28900	9180
	65	172	4225	29584	11180
	78	183	6084	33489	14274
	63	165	3969	27225	10395
	48	162	2304	26244	7776
	53	168	2809	28224	8904
	76	178	5776	31684	13528
	65	170	4225	28900	11050
	74	182	5476	33124	13468
Total	1293	3423	86137	587297	222860

$$\bar{x} = \frac{1293}{20} = 64.65 \qquad \bar{y} = \frac{3423}{20} = 171.15$$

$$\sigma_x = \sqrt{\frac{86137}{20} - 64.65^2} = 11.2795$$

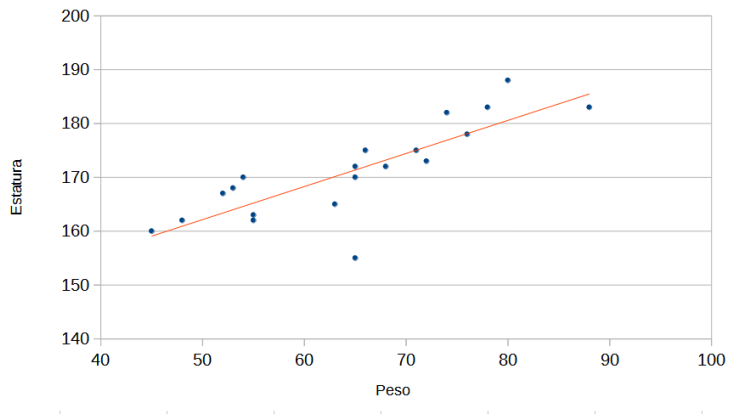
$$\sigma_y = \sqrt{\frac{587297}{20} - 171.15^2} = 8.5163$$

$$\sigma_{xy} = \frac{222860}{20} - 64.65 \cdot 171.15 = 78.1525$$

$$r = \frac{78.1525}{11.2795 \cdot 8.5163} = 0.8136$$

Recta de regresión

Nuestro objetivo es encontrar la recta a la que mejor se adapte nuestra nube de puntos



Sabemos que la ecuación de la recta ha de ser de la forma y = m x + n

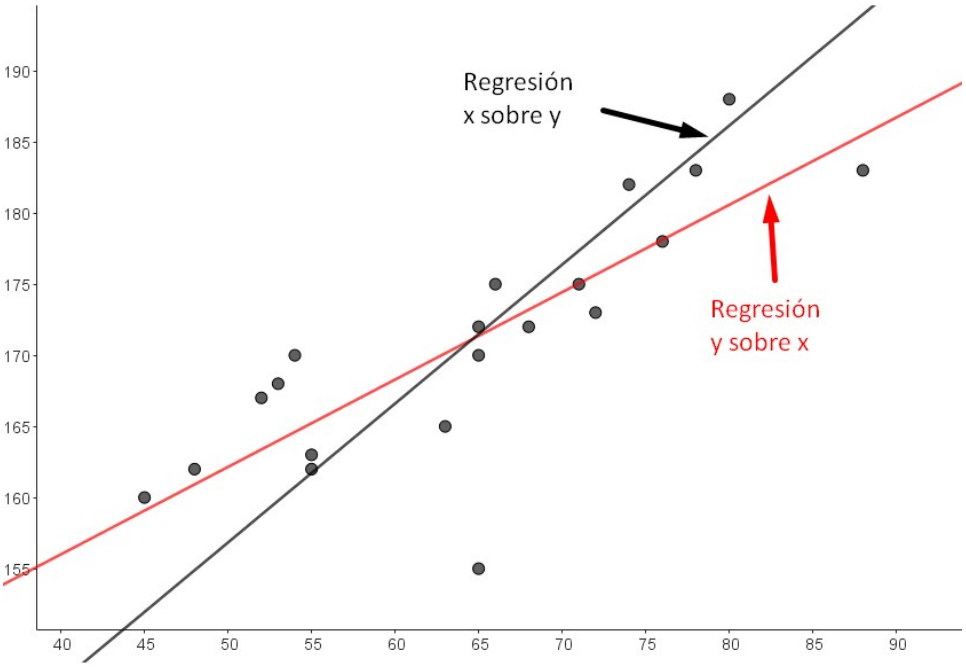
Si la correlación fuese perfecta (coeficiente de correlación igual a uno), entonces para cada punto se tendría que cumplir $y_i = m\,x_i + n$.

Si la correlación no es perfecta, entonces, por lo general $e_i = y_i - (m\,x_i + n) \neq 0$ (a e_i se le llama residuo).

El problema de la recta de regresión consiste en obtener los valores de m y n para los cuales la cantidad $\sum e_i^2$ sea mínima. La solución a este problema es $m = \frac{\sigma_{xy}}{\sigma_x^2}$ y $n = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2} \cdot \bar{x}$

Así pues, observamos:

- La recta de regresión tiene pendiente $m = \frac{\sigma_{xy}}{\sigma_x^2}$
- La recta de regresión pasa por el centro de gravedad $(\bar{x}, \bar{y})$
- Su ecuación punto-pendiente será: $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2} \cdot (x - \bar{x})$
- Hay dos rectas de regresión distintas en función de cual consideremos como variable dependiente (la recta de x sobre y será $x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2} \cdot (y - \bar{y})$). Estas dos rectas solo coinciden si |r|=1 y habrá mayor diferencia entre ellas cuanto menor sea |r|.



Distribuciones condicionadas

Consideremos la siguiente tabla de datos correspondiente a los datos de mortalidad en accidentes de tráfico en la provincia de Pontevedra. Nos interesa estudiar la evolución a lo largo de los años y ver si depende o no del sexo.

Año	Hombres	Mujeres	Total
2008	65	21	86
2009	61	16	77
2010	64	8	72
2011	40	24	64
2012	41	20	61
2013	25	7	32
2014	33	11	44
2015	38	13	51
2016	32	11	43
2017	26	15	41
2018	28	14	42
Total	453	160	613

En este caso tenemos las variables X: “muertos por año” e Y: “muertos según el sexo”.
Sus distribuciones marginales son:

X – Año	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
N.º muertos	86	77	72	64	61	32	44	51	43	41	42	613

Y- sexo	Hombres	Mujeres	Total
N.º muertos	453	160	613

Si hablamos de la distribución de accidentes, condicionada por “ser mujer”, dicha distribución condicionada será:

Año	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
N.º Mujeres muertas	21	16	8	24	20	7	11	13	11	15	14	160

Si calculamos las frecuencias relativas de X (dividiendo cada dato entre el total) obtenemos:

X – Año	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
f rel.	0,140	0,126	0,117	0,104	0,100	0,052	0,072	0,083	0,070	0,067	0,069

comparando con las frecuencias relativas de X condicionado por “ser mujer”

Año (mujeres)	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
f rel.	0,131	0,100	0,050	0,150	0,125	0,044	0,069	0,081	0,069	0,094	0,088

Como las frecuencias relativas de X y de X/“ser mujer” no coinciden, entonces podemos concluir que la variación anual de la mortalidad no es independiente del sexo.