Tema 8: Estadística bidimensional.

8.1 Conceptos de Estadística bidimensional.

- Hasta ahora hemos estudiado una sola variable. Sin embargo, en muchas situaciones interesa analizar dos variables X e Y, al mismo tiempo para estudiar si existe relación entre ellas. Ejemplo: pulso y temperatura de los pacientes de un hospital, los ingresos y gastos de las familias de un colectivo, la edad y los días que faltan al trabajo los trabajadores de una fábrica, etc.
- Tipos de distribuciones bidimensionales:
  - Cualitativa – cualitativa.
  - Cualitativa – cuantitativa (discreta o continua).
  - Cuantitativa (discreta o continua) – cuantitativa (discreta o continua).
- Tipos de tablas:
  - Tabla de dos columnas (x_i , y_i) para pocos datos.
  - Tabla de tres columnas (x_i , y_i , f_i) para muchos datos y pocos valores posibles.
  - Tablas de doble entrada para muchos datos y muchos valores posibles. Ejemplo: las notas de Matemáticas y Física de 20 alumnos:

Notas Mat.	Notas Fís.	Frecuencia
1	2	2
1	3	1
2	3	1
3	2	1
3	5	1
4	3	1
5	1	1
5	2	1
6	1	1
6	2	1
6	5	2
7	6	1
7	7	2
8	2	1
9	8	1
10	9	2
	Total	20

Otro ejemplo: se ha medido el volumen, en litros, y el peso, en kilogramos, de distintos tipos de maletas, obteniéndose los siguientes resultados:

Volumen	97	102	94	107	92	98
Peso	6,9	7,1	6,7	7,4	5,8	6,1

8.2 Cálculo de parámetros bidimensionales.

- Ya conocemos los parámetros unidimensionales para cada variable:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{N} \qquad s_x^2 = \frac{\sum_{i=1}^{n} f_i \cdot (x_i - \overline{x})^2}{N} = \frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{N} - \overline{x}^2$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i \cdot f_i}{N} \qquad s_y^2 = \frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})^2}{N} = \frac{\sum_{i=1}^{n} y_i^2 \cdot f_i}{N} - \overline{y}^2$$

- Ahora aparece un parámetro nuevo: la **covarianza**, que es la media aritmética de las desviaciones de cada una de las variables respecto a sus medias respectivas: $s_{xy} = \frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})(x_i - \overline{x})}{N} = \frac{\sum_{i=1}^{n} f_i \cdot x_i \cdot y_i}{N} - \overline{x} \cdot \overline{y}$
- El **coeficiente de correlación lineal** (coeficiente de Pearson, r) es una forma de cuantificar de forma más precisa el tipo de correlación que hay entre las dos variables: $r = \frac{s_{xy}}{s_x s_y}$

8.3 Correlación o dependencia.

- La **correlación** es la relación o dependencia existente entre las dos variables de una distribución bidimensional. Según sean los diagramas de dispersión, también llamados nube de puntos, podemos establecer los siguientes casos:
  - Independencia funcional o correlación nula: cuando no existe ninguna relación entre las variables ($r = 0$).
  - Dependencia funcional o correlación funcional: cuando existe una función tal que todos los valores de la variable la satisfacen (a cada valor de x le corresponde uno solo de y o a la inversa) ($r = \pm 1$).
  - Dependencia aleatoria o correlación curvilínea (ó lineal): cuando los puntos del diagrama se aproximan a una línea recta o a una curva, puede ser positiva o directa ($0 < r < 1$), o negativa o inversa ($-1 < r < 0$).

8.4 Regresión lineal.

- La **regresión** consiste en ajustar, lo mejor posible, la nube de puntos de un diagrama de dispersión a una curva. Cuando ésta es una recta, se obtiene la recta de regresión lineal. Cuando es una parábola se obtiene una regresión parabólica. Y cuando es una exponencial, se obtiene una regresión exponencial. (Es importante tener en cuenta que en todos los casos, r debe ser distinto de 0).
- Al valor $\frac{s_{xy}}{s_x^2}$ se le llama **coeficiente de regresión de y sobre x** y representa la pendiente de la recta de regresión.

La **recta de regresión de y sobre x** viene dada por: $\boxed{y-\overline{y}=\frac{s_{xy}}{s_x^2}(x-\overline{x})}$.

- De forma análoga, $\boxed{\frac{s_{xy}}{s_y^2}}$ se llama **coeficiente de regresión de x sobre y**. Y la **recta de regresión de x sobre y** es: $\boxed{x-\overline{x}=\frac{s_{xy}}{s_y^2}(y-\overline{y})}$.

En el ejemplo de las notas de Matemáticas y Física.

Notas Mat	Notas Fís	Frecuencia	$x_i \cdot f_i$	$y_i \cdot f_i$	$x_i^2 \cdot f_i$	$y_i^2 \cdot f_i$	$x_i \cdot y_i \cdot f_i$
1	2	2	2	4	2	8	4
1	3	1	1	3	1	9	3
2	3	1	2	3	4	9	6
3	2	1	3	2	9	4	6
3	5	1	3	5	9	25	15
4	3	1	4	3	16	9	12
5	1	1	5	1	25	1	5
5	2	1	5	2	25	4	10
6	1	1	6	1	36	1	6
6	2	1	6	2	36	4	12
6	5	2	12	10	72	50	60
7	6	1	7	6	49	36	42
7	7	2	14	14	98	98	98
8	2	1	8	2	64	4	16
9	8	1	9	8	81	64	72
10	9	2	20	18	200	162	180
	Total	20	107	84	727	488	547

Media x	5,35	media y	4,2
desviación s_x	2,7798381	desviación s_y	2,6
covarianza s_{xy}	4,88	coef. corr. lin. r	0,6751915
Coeficiente de Regresión de y sobre x	0,6315108	Coeficiente de Reg. x sobre y	0,7218935

Las medias son: $\overline{x}=\frac{\sum_{i=1}^{n} x_i \cdot f_i}{N}=\frac{107}{20}=5,35$ e $\overline{y}=\frac{\sum_{i=1}^{n} y_i \cdot f_i}{N}=\frac{84}{20}=4,2$.

Las desviaciones típicas son las siguientes:

$$s_x=\sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (x_i-\overline{x})^2}{N}}=\sqrt{\frac{\sum_{i=1}^{n} f_i \cdot x_i^2}{N}-\overline{x}^2}=\sqrt{\frac{727}{20}-5,35^2}=\sqrt{7,7275}=2,7798381\ .$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 \cdot f_i}{N} - \overline{y}^2} = \sqrt{\frac{488}{20} - 4{,}2^2} = \sqrt{6{,}76} = 2{,}6\,.$$

La covarianza es:

$$s_{xy} = \frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})(x_i - \overline{x})}{N} = \frac{\sum_{i=1}^{n} f_i \cdot x_i \cdot y_i}{N} - \overline{x} \cdot \overline{y} = \frac{547}{20} - 5{,}35 \cdot 4{,}2 = 4{,}88\,.$$

El coeficiente de correlación lineal es: $r = \frac{s_{xy}}{s_x s_y} = \frac{4{,}88}{2{,}779838 \cdot 2{,}6} = 0{,}6751915\,.$

Que indica en este ejemplo que la correlación entre los datos es fuerte.

Los coeficientes de regresión, son los siguientes:

Coeficiente de regresión de y sobre x: $\frac{s_{xy}}{s_x^2} = \frac{4{,}88}{(2{,}7798381)^2} = 0{,}6315108\,.$

Coeficiente de regresión de x sobre y: $\frac{s_{xy}}{s_y^2} = \frac{4{,}88}{(2{,}6)^2} = 0{,}7218935\,.$

Recta de regresión de y sobre x : $y - 4{,}2 = 0{,}6315108 \cdot (x - 5{,}35)$

Recta de regresión de x sobre y: $x - 5{,}35 = 0{,}7218935 \cdot (y - 4{,}2)$

Por ejemplo, si un alumno tiene un 7 en Matemáticas, ¿qué nota se espera que obtenga en Física? La nota de física esperada se obtiene sustituyendo en la recta de regresión de y sobre x:

$y - 4{,}2 = 0{,}6315108 \cdot (7 - 5{,}35) \Rightarrow y = 4{,}2 + 0{,}6315108 \cdot 1{,}65 = 4{,}2 + 1{,}04199282$

Luego se espera que obtenga una nota aproximada de 5,24 en Física.

- En el ejemplo de las maletas: se ha medido el volumen, en litros, y el peso, en kilogramos, de distintos tipos de maletas, obteniendo los resultados que se recogen en esta tabla:

Volumen	97	102	94	107	92	98
Peso	6,9	7,1	6,7	7,4	5,8	6,1

Vol.	Peso	Frecuencia	$x_i \cdot f_i$	$y_i \cdot f_i$	$x_i^2 \cdot f_i$	$y_i^2 \cdot f_i$	$x_i \cdot y_i \cdot f_i$
97	6,9	1	97	6,9	9409	47,61	669,3
102	7,1	1	102	7,1	10404	50,41	724,2
94	6,7	1	94	6,7	8836	44,89	629,8
107	7,4	1	107	7,4	11449	54,76	791,8
92	5,8	1	92	5,8	8464	33,64	533,6
98	6,1	1	98	6,1	9604	37,21	597,8
	Total	6	590	40	58166	268,52	3946,5

media x	98,333333	media y	6,666666
desviación s_x	4,988876	desviación s_y	0,555777
covarianza	2,194444	coef. corr. lin.	0,791445
coef. Regres. de y sobre x	0,088169	coef. Regresión de x sobre y	7,104316

Las medias son: $\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{N} = \frac{590}{6} = 98{,}333333$ $\overline{y} = \frac{\sum_{i=1}^{n} y_i \cdot f_i}{N} = \frac{40}{6} = 6{,}666666$.

Las desviaciones típicas son las siguientes:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (x_i - \overline{x})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot x_i^2}{N} - \overline{x}^2} = \sqrt{\frac{58166}{6} - 98{,}333333^2} = 4{,}988876.$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 \cdot f_i}{N} - \overline{y}^2} == \sqrt{\frac{268{,}52}{6} - 6{,}666666^2} = 0{,}555777.$$

La covarianza es:

$$s_{xy} = \frac{\sum_{i=1}^{n} f_i \cdot (y_i - \overline{y})(x_i - \overline{x})}{N} = \frac{\sum_{i=1}^{n} f_i \cdot x_i \cdot y_i}{N} - \overline{x} \cdot \overline{y} = \frac{3946{,}5}{6} - 98{,}333333 \cdot 6{,}666666 = 2{,}194444.$$

El coeficiente de correlación lineal es:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{2{,}194444}{4{,}988876 \cdot 0{,}555777} = 0{,}791445.$$

Este coeficiente indica que la correlación entre los datos es muy fuerte.

Los coeficientes de regresión, son los siguientes:

Coeficiente de y sobre x: $\frac{s_{xy}}{s_x^2} = \frac{2{,}194444}{(4{,}988876)^2} = 0{,}088169$.

Coeficiente de x sobre y: $\frac{s_{xy}}{s_y^2} = \frac{2{,}194444}{(0{,}555777)^2} = 7{,}104316$.

Recta de regresión de y sobre x : $y - 6{,}666666 = 0{,}088169 \cdot (x - 98{,}333333)$

Recta de regresión de x sobre y: $x - 98{,}333333 = 7{,}104316 \cdot (y - 6{,}666666)$

Por ejemplo, si una maleta tiene un volumen de 120 litros, ¿qué peso se espera que tenga?

El peso esperado se obtiene sustituyendo en la recta de regresión de y sobre x:

$y - 6{,}666666 = 0{,}088169 \cdot (120 - 98{,}333333)$

$y - 6{,}666666 = 0{,}088169 \cdot 21{,}666667$

$y - 6{,}666666 = 1{,}91032827$

$y = 6{,}666666 + 1{,}91032827$

$y = 8{,}57699427$

Luego, se espera que la maleta pese aproximadamente 8.577 gramos.