VARIABLES ESTADÍSTICAS BIDIMENSIONALES.

CONTENIDOS:

- Organización de datos: tablas de frecuencias de doble entrada. Frecuencias marginales.
- Diagrama de dispersión.
- Regresión lineal: rectas de regresión. Coeficiente de correlación lineal. Interpretación.
- Predicción lineal.

Organización de datos.

Las variables estadísticas bidimensionales las representaremos por el par (X,Y), donde X es una variable unidimensional que toma los valores $x_1, x_2, \ldots x_n$ e Y es otra variable unidimensional que toma los valores $y_1, y_2, \ldots y_n$. Si representamos estos pares (x_1,y_1), (x_2,y_2)......en un sistema de ejes cartesianos se obtiene un conjunto de puntos sobre el plano que se denomina **diagrama de dispersión o nube de puntos.**
Los datos pueden ser presentados en dos tipos de tablas: **tabla simple y tabla de doble entrada.** Esta última tabla se puede transformar en una tabla simple.

Cálculo de parámetro.

Consideremos una variable estadística bidimensional (X,Y) y recordemos las definiciones de media y varianza para distribuciones de variable estadística unidimensional:

	Variable X	Variable Y
MEDIA	$\bar{x} = \dfrac{\sum_{i=1}^{p} x_i n_i}{n}$	$\bar{y} = \dfrac{\sum_{i=1}^{p} y_i n_i}{n}$
VARIANZA	$s_x^{\ 2} = \dfrac{\sum_{i=1} n_i x_i^2}{n} - \bar{x}^2$	$s_y^{\ 2} = \dfrac{\sum_{i=1} n_i y_i^2}{n} - \bar{y}^2$

A la raíz cuadrada positiva de las varianzas se la llama **desviación típica** y se representa por s_x y por s_y.
Se llama **covarianza** de una variable bidimensional (X,Y) a la media aritmética de los productos de las desviaciones de cada variable respecto a sus medias respectivas.
La covarianza se representa por s_{xy}

$$s_{xy} = \frac{\sum_{i=1}^{p} n_i x_i\, y_i}{n} - \overline{x}\,\overline{y}$$

Correlación.

Se llama correlación a la teoría que trata de estudiar la relación o dependencia que existe entre las dos variables que intervienen en una distribución bidimensional.

- **La correlación es lineal o curvilínea** según que el diagrama de puntos se condense en torno a una línea recta o a una curva.
- **La correlación es positiva o directa** cuando a medida que crece una variable la otra también crece.
- **La correlación es negativa o inversa** cuando a medida que crece una variable la otra decrece.

- **La correlación es nula** cuando no existe ninguna relación entre ambas variables; en este caso los puntos del diagrama están esparcidos al azar, sin formar ninguna línea, y se dice que las variables no están correlacionadas.
- **La correlación es de tipo funcional** si existe una función que satisface todos los valores de la distribución.

El procedimiento más frecuente utilizado para asignar valores a las distintas correlaciones es a partir del coeficiente de correlación de Pearson. Dicho coeficiente se define mediante la siguiente expresión:

$$r = \frac{s_{xy}}{s_x s_y}$$

El signo del coeficiente r viene dado por el signo de la covarianza, ya que las desviaciones típicas son siempre positivas. El signo de la covarianza decide el comportamiento de la correlación:

- Si la covarianza es positiva, la correlación es directa.
- Si la covarianza es negativa, la correlación es inversa.
- Si la covarianza es nula, no existe correlación.

Veamos que tipo de dependencia existe entre las variables X e Y según el valor de r

1. Si r=-1, todos los valores de la variable (X,Y) se encuentran situados sobre una recta; en consecuencia, satisfacen la ecuación de una recta. Entonces se dice que entre las variables X e Y existe una **dependencia funcional.**
2. Si –1<r<0, la correlación es negativa y será tanto más fuerte a medida que r se aproxime a –1, y tanto más débil a medida que se aproxima a 0. En este caso se dice que las variables X e Y están en **dependencia aleatoria.**
3. Si r=0 no existe ningún tipo de relación entre las dos variables. En este caso se dice que las variables son **aleatoriamente independientes**
4. Si 0<r<1, la correlación es positiva y será tanto más fuerte a medida que r se aproxime a 1, y tanto más débil a medida que se aproxima a 0. En este caso se dice que las variables X e Y están en **dependencia aleatoria**
5. Si r=1, todos los valores de la variable (X,Y) se encuentran situados sobre una recta; en consecuencia, satisfacen la ecuación de una recta. Entonces se dice que entre las variables X e Y existe una **dependencia funcional**

<u>Regresión lineal.</u>

Si entre dos variables existe una fuerte correlación, el diagrama de puntos se condensa en torno a una recta. Sea X la variable independiente e Y la variable dependiente de X, entonces el problema consiste en encontrar la ecuación de la recta que mejor se ajuste a la nube de puntos.

La ecuación buscada será de la forma $y - \overline{y} = a(x - \overline{x})$ donde a es el coeficiente de regresión y es igual a:

$$a = \frac{s_{xy}}{s_x^2}$$

Luego la ecuación de la **recta de regresión de y sobre x** es:

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$$

Sustituyendo en esta ecuación los valores de x podemos obtener, con cierta aproximación, los valores esperados para la variable y , que llamamos estimaciones o previsiones.

La ecuación de la **recta de regresión de x sobre y** es:

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

¿Qué fiabilidad podemos conceder a estos cálculos obtenidos a través de las rectas de regresión? Será tanto mejor cuanto mayor sea el coeficiente de correlación lineal en valor absoluto.

Ejemplo

Se han realizado unas pruebas de habilidad (puntúan de 0 a 5) en un grupo de alumnos. Las siguientes puntuaciones corresponden a las obtenidas por seis alumnos en dos de ellas:

1ª Prueba	**5**	**5**	**4**	**3**	**2**	**4**
2ª Prueba	**4**	**3**	**4**	**4**	**3**	**2**

Calcula la covarianza y el coeficiente de correlación. ¿Cómo es la relación entre las variables?

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
5	4	25	16	20
5	3	25	9	15
4	4	16	16	16
3	4	9	16	12
2	3	4	9	6
4	2	16	4	8
23	20	95	70	77

- Medias:

$$\bar{x} = \frac{23}{6} = 3{,}83$$

$$\bar{y} = \frac{20}{6} = 3{,}33$$

- Desviaciones típicas:

$$\sigma_x = \sqrt{\frac{95}{6} - 3{,}83^2} = \sqrt{1{,}16} = 1{,}08$$

$$\sigma_y = \sqrt{\frac{70}{6} - 3{,}33^2} = \sqrt{0{,}58} = 0{,}76$$

- Covarianza:

$$\sigma_{xy} = \frac{77}{6} - 3{,}83 \cdot 3{,}33 = 0{,}079 \quad \rightarrow \sigma_{xy} = 0{,}079$$

- Coeficiente de correlación:

$$r = \frac{0{,}079}{1{,}08 \cdot 0{,}76} = 0{,}096 \quad \rightarrow r = 0{,}096$$

- La relación entre las variables es prácticamente nula.

Ejemplo-

En seis institutos de la misma zona se ha estudiado la nota media de los estudiantes de 1º de bachillerato en Matemáticas y en Inglés, obteniéndose la información que se recoge en la siguiente tabla:

***X:* Matemáticas**	**6,5**	**5,2**	**6**	**6,5**	**7**	**6**
***Y:* Inglés**	**7**	**5**	**5**	**6**	**7,5**	**5**

a) Halla la recta de regresión de y sobre x.

b)Calcula $\hat{y}(5{,}5)$. ¿Es fiable esta estimación? (Sabemos que $r = 0{,}87$).

a)

x_i	y_i	x_i^2	$x_i y_i$
6,5	7	42,25	45,5
5,2	5	27,04	26
6,0	5	36	30
6,5	6	42,25	39
7	7,5	49	52,5
6	5	36	30
37,2	35,5	232,54	223

- Medias:

$$\bar{x} = \frac{37,2}{6} = 6,2$$

$$\bar{y} = \frac{35,5}{6} = 5,92$$

- Varianza de x:+

$$\sigma_x^2 = \frac{232,54}{6} - 6,2^2 = 0,32$$

- Covarianza:

$$\sigma_{xy} = \frac{223}{6} - 6,2 \cdot 5,92 = 0,46$$

- Coeficiente de regresión:

$$m_{yx} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{0,46}{0,32} = 1,44$$

- Ecuación de la recta de regresión de y sobre x:

$$y = 5,92 + 1,44(x - 6,2) \quad \rightarrow \quad y = 1,44x - 3$$

b) $\hat{y}(5,5) = 1,44 \cdot 5,5 - 3 = 4,92$

Sí es fiable la estimación, puesto que la correlación es fuerte, $r = 0,87$, y $x = 5,5$ está dentro del intervalo de valores que estamos considerando. Por tanto, estimamos que si la nota de Matemáticas es 5,5, la de Inglés será muy probablemente 4,9.

Ejemplo-

Se ha preguntado en seis familias por el número de hijos y el número medio de días que suelen ir al cine cada mes. Las respuestas han sido las siguientes:

***X:* Hijos**	**2**	**1**	**3**	**4**	**2**	**3**
***Y:* Días cine**	**3**	**4**	**4**	**2**	**1**	**4**

a) Halla las dos rectas de regresión y represéntalas.

b) Observando el grado de proximidad entre las dos rectas, ¿cómo crees que será la correlación entre las dos variables?

a)

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
2	3	4	9	6
1	4	1	16	4
3	4	9	16	12
4	2	16	4	8
2	1	4	1	2
3	4	9	16	12
15	18	43	62	44

- Medias:

$$\bar{x} = \frac{15}{6} = 2{,}5$$

$$\bar{y} = \frac{18}{6} = 3$$

- Desviaciones típicas:

$$\sigma_x = \sqrt{\frac{43}{6} - 2{,}5^2} = \sqrt{0{,}92} = 0{,}96$$

$$\sigma_y = \sqrt{\frac{62}{6} - 3^2} = \sqrt{1{,}33} = 1{,}15$$

- Covarianza:

$$\sigma_{xy} = \frac{44}{6} - 2{,}5 \cdot 3 = -0{,}17$$

- Coeficientes de regresión:

y sobre x $\rightarrow$ $m_{yx} = \dfrac{-0{,}17}{0{,}92} = -0{,}18$

x sobre y $\rightarrow$ $m_{xy} = \dfrac{-0{,}17}{1{,}33} = -0{,}13$

- Rectas de regresión:

y sobre x $\rightarrow$ $y = 3 - 0{,}18(x - 2{,}5)$ $\rightarrow$ $y = -0{,}18x + 3{,}45$

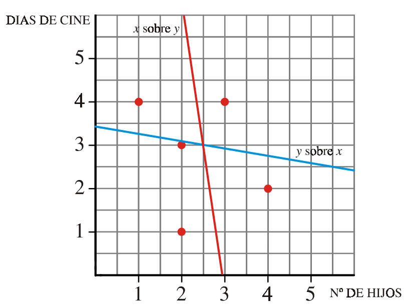
x sobre y $\rightarrow$ $x = 2{,}5 - 0{,}13(y - 3)$

$$x = -0{,}13y + 2{,}89$$

$$0{,}13y = 2{,}89 - x$$

$$y = \frac{-x + 2{,}89}{0{,}13} \rightarrow y = -7{,}69x + 22{,}23$$

- Representación:



b) La correlación es prácticamente nula; las rectas son casi perpendiculares.

EJERCICIOS.

1. La siguiente tabla ofrece los resultados de 6 pares de observaciones, realizadas para analizar el grado de relación existente entre dos variables X e Y:

X	2	2	3	3	3	4
Y	0	1	1	2	4	3

Obtener:
a) Recta de regresión de Y sobre X.
b) Representación gráfica de la misma, así como de los pares de observaciones anteriores.
c) ¿Qué grado de relación lineal existe entre ambas variables?

2. Dada esta distribución bidimensional:

X	5	6,5	8	4	3
Y	4,5	7	7,5	5	3,5

a) Calcular el coeficiente de correlación lineal, interpretando el resultado.
b) Determinar la recta de regresión de Y sobre X.
c) Hallar el punto donde se cortan las dos rectas.

3. Cinco niñas de 2,3,5,7 y 8 años de edad pesan, respectivamente, 14,20,32,42 y 44 kilos.
a) Hallar la ecuación de la recta de regresión de la edad sobre el peso.
b) ¿Cuál sería el peso aproximado de una niña de 6 años?

4. Las notas obtenidas por cinco alumnos en Matemáticas y Música son:

Matemáticas	6	4	8	5	3,5
Música	6,5	4,5	7	5	4

Determinar las rectas de regresión y calcular la nota esperada en Música para un alumno que tiene 7,5 en Matemáticas.

5. La media de los pesos de una población es de 65 kg y la de las estaturas 170 cm, mientras que las desviaciones típicas son de 5 kg y 10 cm, respectivamente, y la covarianza de ambas variables es 40. Calcular la recta de regresión de los pesos respecto de las estaturas ¿Cuánto se estima que pesará un individuo de 180 cm de estatura?

6. La tabla siguiente nos da las notas del test de aptitud (X) dadas a 6 dependientes a prueba y ventas del primer mes de prueba (Y) en cientos de pesetas:

X	25	42	33	54	29	36
Y	42	72	50	90	45	48

a) Hallar el coeficiente de correlación e interpretar el resultado obtenido.
b) Hallar la recta de regresión de Y sobre X. Predecir las ventas de un vendedor que obtenga 47 en el test.

7. Se ha observado una variable estadística bidimensional y se ha obtenido la siguiente tabla:

		X		
		100	50	25
Y	14	1	1	-
	18	2	3	-
	22	-	1	2

Se pide:
a) Calcular la covarianza.
b) Obtener e interpretar el coeficiente de correlación lineal.

c) Ecuación de la recta de regresión Y sobre X.

8. Los valores de dos variables X e Y se distribuyen según la tabla siguiente. Determinar el coeficiente de correlación y la recta de regresión Y sobre X. Comentar lo fiables que son las predicciones basadas en esa recta.

		X		
		0	2	4
Y	1	2	1	3
	2	1	4	2
	3	2	5	0

9. Las puntuaciones obtenidas por un grupo de alumnos de COU en una batería de test que mide la habilidad verbal X y el razonamiento abstracto Y son las siguientes:

		X			
		20	30	40	50
Y	(25,35)	6	4	-	-
	(35,45)	3	6	1	-
	(45,55)	-	2	5	3
	(55,65)	-	1	2	7

Se pide:

a) ¿Existe correlación entre ambas variables?

b) Según los datos de la tabla, si uno de estos alumnos obtiene una puntuación de 70 puntos en razonamiento abstracto, ¿en cuánto se estimará su habilidad verbal?.

10. Los valores de dos variables aleatorias X e Y se distribuyen según la tabla:

X	1	1	1	2	3	3	3
Y	0	2	4	2	2	2	3
n_i	2	1	3	4	2	5	3

a) Determina el coeficiente de correlación y la recta de regresión de Y sobre X

b) Comenta la fiabilidad de las predicciones basadas en esa recta.