Tema 9: ESTADÍSTICA BIDIMENSIONAL.

9.1 Nubes de puntos. Correlación.

"Regresión" es un término acuñado por sir Francis Galton (1822-1911) al comparar la altura de padres e hijos. Al lanzar una piedra hacia arriba, hay una **fórmula** que nos permite calcular, exactamente, la altura alcanzada en función de la velocidad con que se lanza. Se dice que es una **relación funcional**. El peso y la estatura de las personas están muy relacionadas (correlación) pero no se podría dar una fórmula que permita conocer el peso exacto de una persona sabiendo su estatura. Se dice que es una **relación estadística**. Si tenemos un colectivo de n individuos y estudiamos dos variables x e y, al conjunto de pares de valores (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) se llama **distribución bidimensional** (a cada individuo le corresponden los valores de dos variables). Tomando esos dos valores como las coordenadas de un punto, dicha distribución se puede representar obteniendo una **nube de puntos** o **diagrama de dispersión**.

Ejemplo 1: Recogemos las notas obtenidas en el 2º trimestre en las materias de Economía, Lengua y Matemáticas en 1º de Bachillerato C y lo mostramos en la siguiente **tabla simple**:

Matemáticas	4	7	3	6	6	8	3	7	2	7	3	6	3	8	6
Economía	8	10	8	9	8	10	8	8	5	8	4	8	8	10	7
Lengua	5	9	5	7	4	9	5	8	3	9	4	8	6	7	7

Matemáticas	7	4	4	1	2	5	2	2	1	4	5	8	7	6
Economía	9	8	9	4	7	9	5	8	4	9	7	9	9	7
Lengua	9	4	5	3	4	8	4	6	3	4	7	8	5	8

¿Cuál es la población del estudio? ¿Qué variables se estudian? ¿De qué tipo son dichas variables?

Dichos datos pueden ser mostrados en otro tipo de tablas, **tabla de contingencia.** Anota en cada celda el número de individuos que verifican dichas variables, siendo Economía (X) y Matemáticas (Y):

	1	2	3	4	5	6	7	8	9	10	Y
1											
2											
3											
4											
5											
6											
7											
8											
9											
10											
X											

Dichos datos se pueden representar en los ejes de coordenadas. Una de las variables se indicaran en el eje x y la otra en el eje y, cada pareja (x,y) forma un punto y con todas las parejas formamos la **nube de puntos** o **diagrama de dispersión**.

Tenemos un colectivo de n individuos, si estudiamos dos variables x e y, al conjunto de pares de valores (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) se llama **distribución bidimensional** (a cada individuo le corresponden los valores de dos variables). Tomando esos dos valores como las coordenadas de un punto, dicha distribución se puede representar obteniendo una **nube de puntos** o **diagrama de dispersión**. ¿Están relacionadas ambas variables? Para responder a dicha pregunta debemos conocer el concepto de correlación. Si la respuesta es afirmativa, calcularemos una recta que nos permite calcular una variable en función de la otra, esa recta se llama **recta de regresión**.

Ejercicio 1: Representa la nube de puntos de:

a) Matemáticas – Economía. Con la tabla de contingencia.
b) Matemáticas – Lengua.
c) Economía – Lengua.

Ejemplo 2: La siguiente tabla muestra la puntuación (Y) obtenida por 1000 personas en función de su edad (X) en un test de aritmética en una tabla de doble entrada:

X \ Y	(125, 175]	(175, 225]	(225, 275]	(275, 325]	(325, 375]
(15, 35]	23	62	163	94	28
(35, 55]	24	55	159	80	22
(55, 75]	33	65	127	53	12

El estudio de la distribución conjunta de una variable dimensional (X, Y) implica analizar cada una de las variables por medio de sus distribuciones marginales:

9.2 Medida de la correlación.

Cuando trabajamos con distribuciones bidimensionales, debemos comprobar si la relación entre ambas variables es fuerte o débil. Para ello vamos a distintos conceptos a partir de las variables marginales y conjuntas:

a) **Centro de gravedad** que es el punto $(\bar{x}, \bar{y})$, siendo $\bar{x} = \frac{\sum x_i}{n}$, $\bar{y} = \frac{\sum y_i}{n}$.
b) **Covarianza** que viene dada por una doble fórmula, siendo la segunda la más cómoda
$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y}$$
El valor de la covarianza depende de las unidades de medida de las variables X e Y. Su signo informa acerca de la variación relativa de las variables, si es positiva a mayor X mayor Y, si es negativa, a mayor X, menor Y.
c) Se define como **coeficiente de correlación muestral** o **Correlación** al valor que viene dado por la expresión $r = \frac{s_{xy}}{s_x s_y}$.

Observaciones sobre la correlación:

1) La correlación no tiene unidades, por ello podemos comparar estudios con distintas unidades.
2) Si r = 1 o r = -1, se dice que la correlación es perfecta.
3) La correlación es fuerte si el valor absoluto de r es cercano a 1.
4) La correlación es débil si el valor absoluto de r es cercano a 0.
5) Si r es positivo la relación es directa (es decir, a mayor X, mayor Y) y si r es negativo, la relación es inversa (a mayor X, menor Y).

Ejercicio 2: Completa las tablas con las distribuciones marginales de los ejemplos del archivo "Ejercicios de clase de la UD 9: Estadística Bidimensional" y realiza las nubes de puntos correspondientes, excepto en el ejemplo 4.

Ejercicio 3: Calcula la correlación de los ejemplos del archivo anterior.

Si la **correlación** tiene un valor próximo a uno, las variables son dependientes, conociendo X podemos calcular Y, o viceversa, siempre que el valor que queremos predecir esté dentro del rango de los valores de la muestra. Para estos valores tiene sentido hablar de la **recta de regresión**. Al representar la recta de regresión veremos que la nube de puntos está muy cerca de la recta de regresión. En cambio, si r está cerca de 0, las variables son independientes.

9.3 Recta de regresión.

Si la correlación entre ambas variables es fuerte, podremos calcular la recta que las relaciona cuya ecuación se consigue estudiando que la distancia de la recta a la nube de puntos sea mínima (cuadrado de las distancias mínimo). Para construir la **recta de regresión de Y sobre X** debemos tener en cuenta $m_{yx} = \frac{s_{xy}}{s_x^2}$ llamado **coeficiente de regresión** y pasa por el punto $(\bar{x}, \bar{y})$, por tanto la recta queda de la forma: $y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x})$

Con ella podemos estimar el valor de y dado el de x, con x dentro de los valores del rango de x_i. Lo denotamos por $\hat{y}(x_i)$, es decir, $\hat{y}(x_i)$ es el valor aproximado de y para x_i.

Si nosotros queremos saber el de x conocido el valor de y, debemos utilizar la **recta de regresión de X sobre Y**, que se construye de forma análoga a la anterior: $x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y})$

Si la correlación es casi nula, las dos rectas serán casi perpendiculares, si la correlación es fuerte, el ángulo que forman es pequeño y la nube de puntos está muy cercana a ambas, si la r es casi 1, las rectas son prácticamente coincidentes.

Ejercicio 4: Calcula la recta de regresión de Y sobre X y de X sobre Y de los seis ejemplos.