MATEMÁTICA Y ESTADÍSTICA

Unidad 5

Análisis descriptivo de dos variables conjuntas.

Tablas de contingencia, distribuciones marginales y condicionales. Medidas de correlación lineal. Recta de regresión.

Roberto Fiadone

Unidad 5. Distribuciones bidimensionales

Introducción

Cuando realizamos un análisis estadístico y observamos simultáneamente dos características en **un mismo individuo** obtenemos pares de resultados, por ejemplo, al observar en cada persona el color de sus ojos y el color de su pelo.

Los distintos valores que pueden adoptar estos caracteres forman un conjunto de pares que llamamos ***variable bidimensional.*** Así, si intentamos relacionar el peso y la altura de las personas, a cada persona, le asociamos un par de valores: (peso, altura).

Los pares de valores así formados pueden resumirse en tablas de frecuencias o gráficos en forma similar a las distribuciones de una variable.

En el caso en que las variables que se estudian sean **ambas cualitativas,** o **una cualitativa y la otra cuantitativa discreta**, las tablas de frecuencias reciben el nombre de **tablas de contingencia.** Estas son el resultado del cruce de las dos variables donde se han volcado las frecuencias correspondientes a cada combinación de las variables.

Ejemplo 1

Veinte estudiantes universitarios fueron clasificados por sexo y por orientación de la carrera que cursan (exactas, económicas o humanidades) como se muestra a continuación:

Estudiante	Sexo	Orientación	Estudiante	Sexo	Orientación
1	F	Humanidades	11	M	Económicas
2	M	Exactas	12	F	Económicas
3	M	Económicas	13	F	Exactas
4	F	Humanidades	14	M	Económicas
5	F	Económicas	15	M	Económicas
6	F	Económicas	16	F	Humanidades
7	M	Económicas	17	F	Humanidades
8	M	Exactas	18	F	Exactas
9	M	Económicas	19	M	Exactas
10	M	Humanidades	20	M	Económicas

Estos datos pueden resumirse en una tabla que muestre conjuntamente la información. Cada una de las filas de la tabla representa los valores de la variable “sexo” y en las columnas los valores de las variable “orientación de la carrera**”.**

En cada celda contabilizamos las observaciones que cumplen simultáneamente con las dos características:

	Orientación		
Sexo:	**Humanidades**	**Exactas**	**Económicas**
Masculino	1	3	7
Femenino	4	2	3

Agreguemos ahora a la tabla los totales de filas y columnas:

	Orientación			
Sexo:	**Humanidades**	**Exactas**	**Económicas**	**Total**
Masculino	1	3	7	11
Femenino	4	2	3	9
Total	5	5	10	20

Los casilleros coloreados contienen las llamadas **frecuencias conjuntas.**

Mientras que las frecuencias que aparecen en los márgenes, o sea, los subtotales, son las **frecuencias marginales**.

Si, por ejemplo, nos preguntamos qué porcentaje de los universitarios estudian económicas, la respuesta sería $\frac{10}{20} \cdot 100 = 50\%$

Pero si nos preguntamos, que porcentaje de las **universitarias mujeres estudian económicas**, nos centramos solamente en la fila de las mujeres, como si el resto de la tabla no existiera…

	Orientación			
Sexo:	**Humanidades**	**Exactas**	**Económicas**	**Total**
Femenino	4	2	3	9

Y calculamos el porcentaje sobre el total de mujeres: $\frac{3}{9} \cdot 100 = 33,\hat{3}\%$

Si en cambio nos preguntamos qué porcentaje de los que **cursan económicas son mujeres**, nos centraríamos solamente en aquellos que cursan económicas…

Sexo:	**Económicas**
Masculino	7
Femenino	3
	10

Y concluiríamos que el $\frac{3}{10} \cdot 100 = 30\%$ son mujeres.

¿Qué porcentaje de los que estudian humanidades o exactas son mujeres? Para responder a esto nos centramos solamente en los que estudian humanidades o exactas…

	Orientación	
Sexo:	**Humanidades**	**Exactas**
Masculino	1	3
Femenino	4	2
Total	5	5

Contamos ahora cuantas mujeres son las que cursan humanidades o exactas:

$$4 + 2 = 6$$

Y cuantos universitarios en total (hombre o mujer) cursan humanidades o exactas:

$$5 + 5 = 10$$

Por lo tanto, de los que cursan solo humanidades o exactas, $\frac{6}{10} \cdot 100 = \mathbf{40}\%$ son mujeres.

Ejemplo 2

La siguiente tabla resume los resultados de una observación a un grupo de niños, clasificados por el color de ojos y el color de cabello.

	Color de cabello		
Color de ojos	**Rubio**	**Moreno**	**Total**
Marrones	25	20	**45**
Azules	15	15	**30**
Verdes	40	35	**75**
Total	**80**	**70**	**150**

Las variables que intervienen son ambas cualitativas.

Podemos hacer las siguientes observaciones respecto a los niños, por ejemplo:

- El porcentaje de niños con cabello rubio y ojos marrones es de $\frac{25}{150} \cdot 100 = \mathbf{16,67}\%$ Tiene cabello rubio y ojos marrones y un 10% cabello moreno y ojos azules.
- $\frac{80}{150} \cdot 100 = \mathbf{53,33}\%$ son rubios y un $\frac{75}{150} \cdot 100 = 50\%$ tiene ojos verdes.

Si ahora nos interesa saber *qué porcentaje de niños **rubios** tiene los ojos azules,* vemos que, **de los 80 niños rubios, hay 15 que tienen ojos azules**, calculamos:

$$\frac{15 \cdot 100}{80} = \mathbf{18,75}\%$$

Es decir, nos **restringimos** a ver, dentro de la columna de rubios, que porcentaje son de ojos azules, *como si las restantes columnas no existiesen*.

Realizando lo mismo para los demás colores de ojos obtendríamos que la distribución del color de ojos *condicionada al color de cabello rubio* es la siguiente:

Color de	**Frecuencia**	**Porcentaje**
Marrones	25	31,25%
Azules	15	18,75%
Verdes	40	50%
Total	**80**	**100%**

Dos variables cuantitativas, recta de regresión

Vamos a realizar ahora el análisis de observaciones bivariadas, donde **ambas variables son cuantitativas.**

Ejemplo 3

Supongamos que se evaluó a 10 estudiantes en lengua e inglés con puntuación de 0 a 100, registrándose la información en la siguiente tabla.

Lengua:	30	24	82	48	64	60	73	96	18	54
Inglés:	37	30	52	52	60	66	61	71	22	44

La tabla resume distribuciones de dos variables cuantitativas:

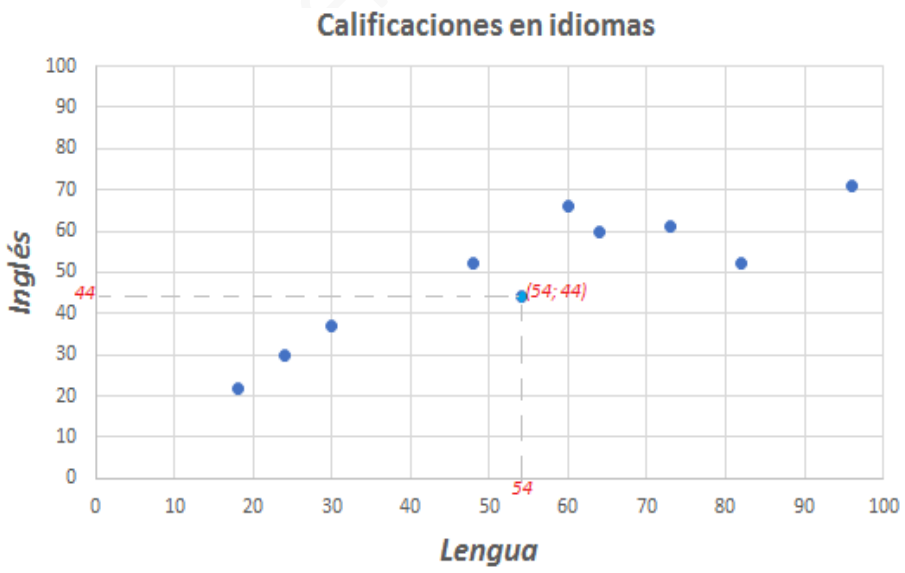
- Notas de los estudiantes en lengua
- Notas de los estudiantes en inglés

Nos interesa saber si ambos conjuntos de datos están relacionados, es decir, si la nota que obtiene un estudiante en una materia está relacionada con la que obtuvo en la otra, de manera que podamos anticipar cuál será aproximadamente su nota en una materia conociendo la nota de la otra, o si en realidad no hay ninguna relación entre ellas.

Para una primera aproximación a este análisis, utilizaremos lo que se denomina un gráfico de dispersión.

Gráfico de dispersión

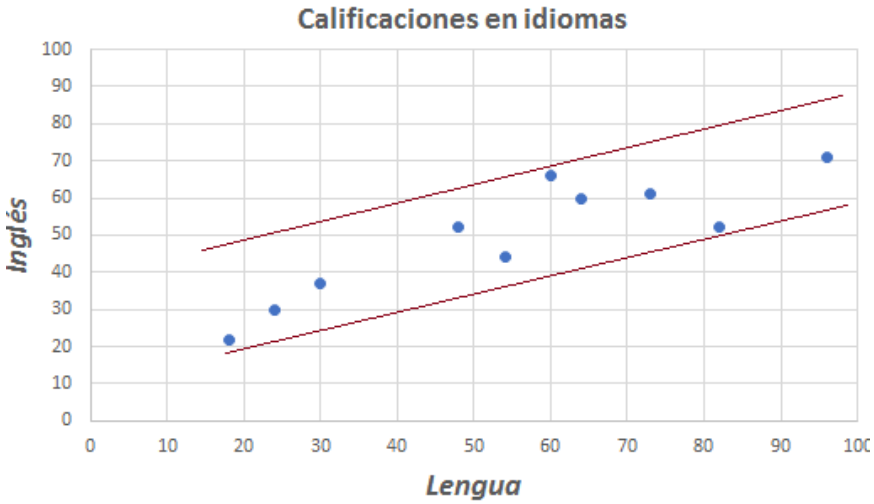
Comencemos por graficar los datos. Para ello, usaremos un tipo de gráfico llamado ***diagrama*** o ***gráfico de dispersión*** o también, ***nube de puntos***.



En el diagrama, cada uno de los 10 alumnos está representado por un punto cuya coordenada sobre el eje x representa la nota en lengua del alumnos, y sobre el eje y la nota en inglés. Así, por ejemplo, el punto de coordenadas $(54;\ 44)$ está representando al alumnos que en Lengua obtuvo 54 puntos y en inglés 44.

Notemos además que, si bien los puntos no se encuentran perfectamente alineados sobre una recta, están ubicados sobre una franja que va desde el extremo inferior

izquierdo hacia el extremo superior derecho, que nos da una idea de cierta tendencia*: salvo pocas excepciones, los estudiantes con bajas notas en lengua suelen tener bajas notas en inglés, y los que obtienen altas notas en lengua, también tienen altas notas en inglés.*

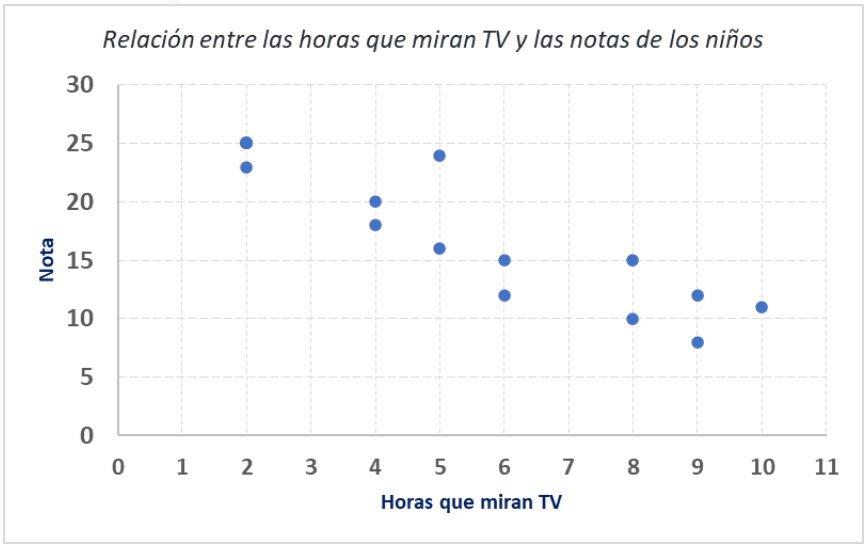


Ejemplo 4

Consideramos ahora otro grupo de datos que resume las observaciones relativas a un grupo de niños/as, en dónde se estudia la cantidad de horas que dedican a ver televisión diariamente y las calificaciones promedio (de 1 a 100) obtenidas al finalizar el cuatrimestre.

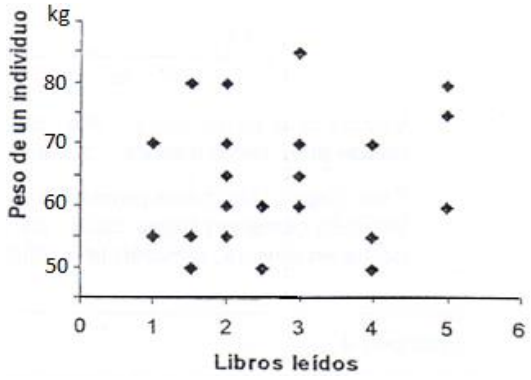
Horas que miran TV	5	2	6	8	5	10	4	6	9	2	4	8	9
Notas	16	25	12	10	24	11	20	15	8	23	18	15	12

Si dibujamos el **diagrama de dispersión o nube de puntos**, la disposición en que aparecen los puntos permite suponer que, cuantas más horas los niños/as miran televisión, es menor su rendimiento escolar.



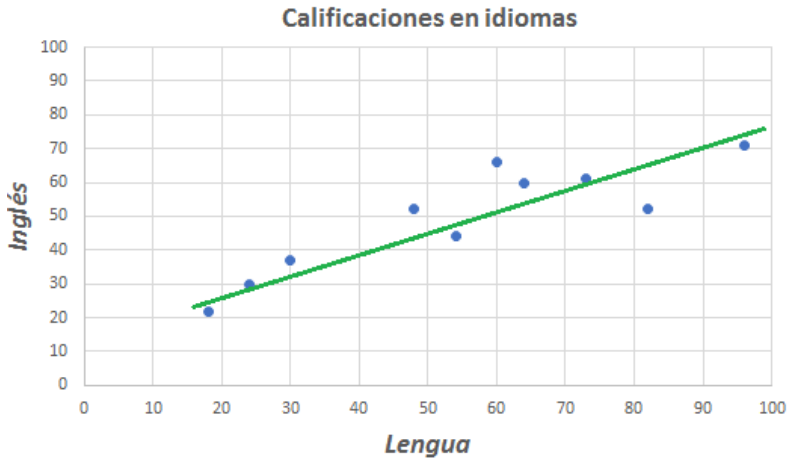
En ambos ejemplos podemos suponer que hay una conexión entre las variables observadas. Es decir, la forma en que una varía está relacionada con la variación de la otra. A esta relación la llamamos **correlación.**

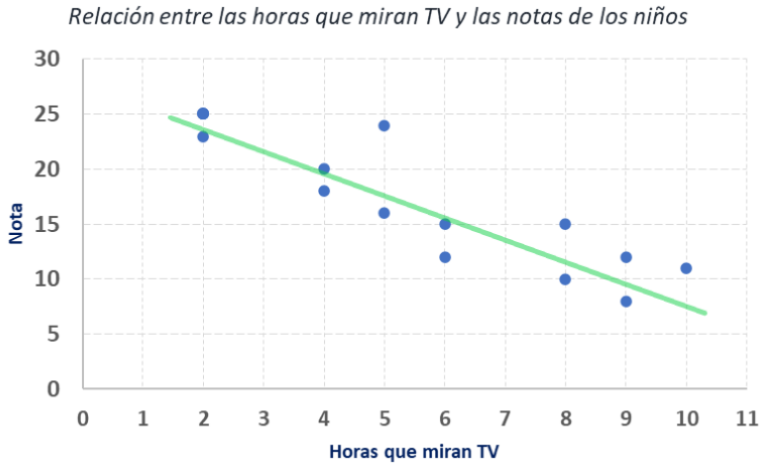
No siempre existe algún tipo de conexión entre las dos variables estudiadas. Así, por ejemplo, el siguiente gráfico muestra, para un grupo de personas elegidas al azar, la cantidad de libros leídos por semana y el peso del mismo individuo.



En este caso (como era de esperarse…) no se ve ninguna relación entre ambas variables.

En definitiva, el diagrama de dispersión nos da una idea de que puede existir algún tipo de relación lineal entre dos variables cuantitativas. A esa supuesta recta a la cual parecen acercarse los puntos que representan cada observación se la llama ***recta de regresión***.





La nube de puntos nos permite apreciar la mayor o menor relación entre las variables (correlación) y, **la pendiente de la recta de regresión, ver si esta correlación es positiva o negativa.**

Medida de la correlación

Vimos que algunas nubes de puntos sugieren rectas. Cuanto más apretados estén los puntos respecto a la recta mayor es la correlación. Nos interesa poder hallar un indicador de la intensidad de la relación entre las variables.

A este indicador lo llamamos ***coeficiente de correlación lineal de Pearson.***

Para calcular este coeficiente necesitamos establecer una medida de la variabilidad conjunta de ambas variables. Esta medida recibe el nombre de ***covarianza.*** Este número mide la variación conjunta de ambas variables.

La fórmula de cálculo es la siguiente.

$$\boldsymbol{Cov_{XY}} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N} = \frac{\boldsymbol{\Sigma x_i y_i}}{\boldsymbol{N}} - \boldsymbol{\bar{x} \cdot \bar{y}}$$

Siendo X e Y las dos variables de las cuales se va a estudiar la variabilidad conjunta o covarianza. **La fórmula de más a la derecha es más simple de aplicar al momento de realizar los cálculos.**

Veamos como aplicamos la fórmula en el primer ejemplo, el de las notas en lengua e inglés. Para ello anexamos una columna con los productos $x_i \cdot y_i$ que necesitaremos para el cálculo de la Cov_{XY}:

Lengua (x_i)	Inglés (y_i)	$x_i \cdot y_i$
30	37	(30·37=) 1110
24	30	(24·30=) 720
82	52	4264
48	52	2496
64	60	3840
60	66	3960
73	61	4453
96	71	6816
18	22	396
54	44	2376
549	**495**	**30.431**

En la última fila figuran los totales, y de ella deducimos que, dado que $N = 10$:

$$\bar{x} = \frac{549}{10} = 54{,}9 \qquad \bar{y} = \frac{495}{10} = 49{,}5$$

$$\boldsymbol{Cov_{XY}} = \frac{\Sigma x_i y_i}{N} - \bar{x} \cdot \bar{y} = \frac{30431}{10} - 54{,}9 \cdot 49{,}5 = \boldsymbol{325{,}5}$$

Para hallar el llamado **coeficiente de correlación lineal**, al que denotaremos por $\boldsymbol{\rho}$, dividimos la covarianza por el producto de los desvíos de ambas distribuciones.

$$\boldsymbol{\rho = \frac{Cov_{xy}}{\sigma_x \cdot \sigma_y}}$$

Realizando los cálculos correspondientes estudiados en la unidad 4 **para el cálculo de los desvíos estándar,** se llega a que**,** para las calificaciones en lengua, es $\sigma_x = 24{,}18$ y para el de las calificaciones en inglés es $\sigma_y = 15{,}2$**.**

Por lo tanto:

$$\boldsymbol{\rho = \frac{325{,}5}{24{,}18 \cdot 15{,}2} = 0{,}88}$$

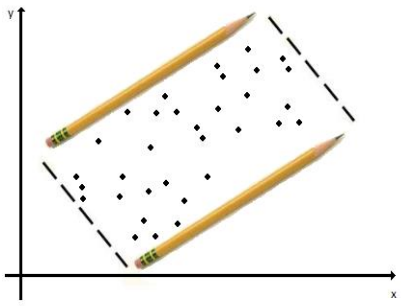
Muy importante:

- El valor de $\boldsymbol{\rho}$ oscilará **siempre** entre $\boldsymbol{-1\ y\ 1, o\ sea,}$ $(\boldsymbol{-1 \leq \rho \leq 1})$.
- **Cuanto más cercano a <u>cero</u> sea el valor de $\boldsymbol{\rho}$, <u>menor es la correlación</u>** entre ambas variables.
- **Cuanto más cercano a <u>uno</u> sea el valor de $\boldsymbol{\rho}$,** la correlación entre las variables **<u>será mayor</u> y al incrementar una de ellas, la otra se incrementará** también.
- **Cuanto más cercano a <u>menos uno</u> sea el valor de $\boldsymbol{\rho}$,** la correlación **será también mayor**, **pero al incrementar una de ellas, la otra disminuirá su valor**.

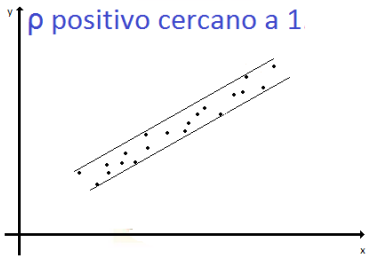
En el ejemplo visto, el coeficiente de correlación lineal fue de $0{,}88$, que es bastante cercano a 1, por lo que podemos concluir que las dos variables están bien correlacionadas y el hecho de que nos haya dado una correlación positiva nos indica que al incrementarse una se incrementará también la otra.

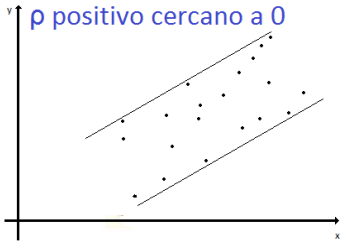
Estimación visual del coeficiente de correlación lineal

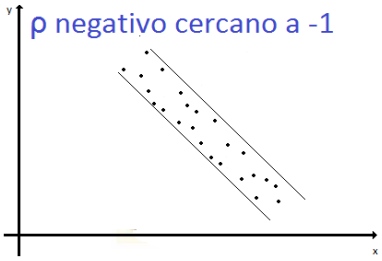
Puede tenerse una idea aproximada de que tan buena, y de que signo será la correlación. Coloque paralelamente dos lápices sobre el diagrama de dispersión y muévalos de modo que estén lo más cerca posible y que todos los puntos del diagrama estén entre ellos. Entonces se puede visualizar una región rectangular encerrada por los dos lápices y que termina de manera justa en los puntos extremos del diagrama de dispersión: cuanto más largo que ancho es el rectángulo creado, mayor es la correlación.

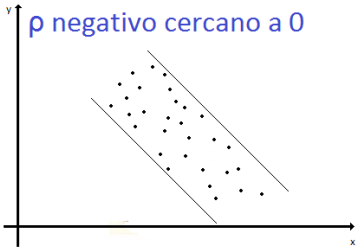


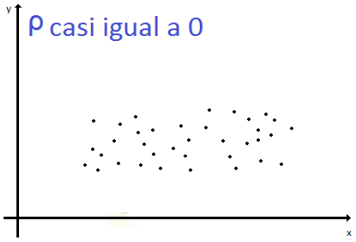
El signo de ρ se determina por la posición general del largo de la región rectangular. Si el largo está en posición creciente, ρ es positivo; si está en posición decreciente, ρ es negativo. Si el rectángulo está en posición horizontal o vertical, entonces ρ es cero, sin importar la razón del largo al ancho.

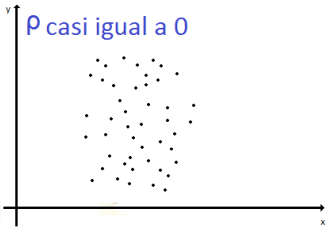

ρ positivo cercano a 1


ρ positivo cercano a 0


ρ negativo cercano a -1


ρ negativo cercano a 0


ρ casi igual a 0


ρ casi igual a 0