Capítulo III	DISTRIBUCIONES BIDIMENSIONALES

3.1. Introducción.

Estudiaremos dos características de un mismo elemento de la población (altura y peso, dos asignaturas, longitud y latitud).
De forma general, si se estudian sobre una misma población y se miden por las mismas unidades estadísticas una variable X y una variable Y, se obtienen series estadísticas de las variables X e Y.
Considerando simultáneamente las dos series, se suele decir que estamos ante una variable estadística bidimensional.

3.2. Tabulación de variables estadísticas bidimensionales.

Vamos a considerar 2 tipos de tabulaciones:
1º) Para variables cuantitativas, que reciben el nombre de tabla de correlación.
2º) Para variables cualitativas, que reciben el nombre de tabla de contingencia.

3.2.1.Tablas de correlación.

Sea una población estudiada simultaneamente según dos caracteres X e Y; que representaremos genéricamente como $\mathbf{(x_i; y_j ; n_{ij})}$, donde x_i; y_j, son dos valores cualesquiera y n_{ij} es la frecuencia absoluta conjunta del valor i-ésimo de X con el j-ésimo de Y.

Una forma de disponer estos resultados es la conocida como tabla de doble entrada o tabla de correlación, la cual podemos representar como sigue:

Y / X	y_1	y_2	.....	y_j	.....	y_s	$n_{i\cdot}$	$f_{i\cdot}$
x_1	n11	n12	.....	n1j	.....	n1k	$n_{1\cdot}$	$f_{1\cdot}$
x_2	n21	n22	.....	n2j	.....	n2k	$n_{2\cdot}$	$f_{2\cdot}$
. . .	. .	. .	. .... .	. .	. .... .	. .	. .	. .
x_i	ni1	ni2	.....	nij	.....	nik	$n_{i\cdot}$	$f_{i\cdot}$
. . .	. . .	. . .	. .... .	. . .	. .... .	. . .	. . .	. . .
x_r	nh1	nh2	.....	nhj	.....	nhk	$n_{h\cdot}$	$f_{h\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	.....	$n_{\cdot j}$	.....	$n_{\cdot k}$	N	
$f_{\cdot j}$	$f_{\cdot 1}$	$f_{\cdot 2}$	.....	$f_{\cdot j}$	.....	$f_{\cdot k}$		1

En este caso, n_{11} nos indica el número de veces que aparece x_1 conjuntamente con y_1; n_{12}, nos indica la frecuencia conjunta de x_1 con y_2, etc.

- **Tipos de distribuciones**

Cuando se estudian conjuntamente dos variables, surgen tres tipo de distribuciones: Distribuciones conjuntas, distribuciones marginales y distribuciones condicionadas.

a) Distribución conjunta

- *La frecuencia absoluta conjunta*, viene determinada por el número de veces que aparece el par ordenado (x_i , y_j), y se representa por " **n_{ij} ".**

- *La frecuencia relativa conjunta*, del par (x_i , y_j) es el cociente entre la frecuencia absoluta conjunta y el número total de observaciones. Se trata de " **f_{ij} ".**

Se cumplen las siguientes relaciones entre las frecuencias de distribución conjunta:
1ª) La suma de las frecuencias absolutas conjuntas, extendida a todos los pares es igual al total de observaciones.

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = N$$

2ª) La suma de todas las frecuencias relativas conjuntas extendida a todos los pares es igual a la unidad.

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = 1$$

b) Distribuciones marginales

Cuando trabajamos con más de una variable y queremos calcular las distribuciones de frecuencias de cada una de manera independiente, nos encontramos con las distribuciones marginales.

Variable X

x_i	$n_i.$	$f_i.$
x_1	$n_1.$	$f_1.$
x_2	$n_2.$	$f_2.$
x_3	n_3 .	f_3 .
x_4	$n_4.$	$f_4.$
	N	1

Variable Y

y_j	$n.j$	$f.j$
y_1	$n.1$	$f.1$
y_2	$n.2$	$f.2$
y_3	$n.3$	$f.3$
y_4	$n.4$	$f.4$
	N	1

- *Frecuencia absoluta marginal: el valor* n_i. Representa el número de veces que aparece el valor x_i de X, sin tener en cuenta cual es el valor de la variable Y. A $n_{i.}$ se le denomina frecuencia absoluta marginal del valor x_i de X, de forma que:

-

$$n_{i.} = n_{i1+}n_{i2} + ... + n_{is}$$

De la misma manera, la frecuencia absoluta marginal del valor y_j de Y se denotará por $n_{.j}$

$$n_{.j} = n_{1j+} n_{2j} + ... + n_{rj}$$

- *Frecuencia relativa marginal*

La frecuencia relativa marginal de x_i de X, viene dada por:

$$f_{i.} = \frac{n_{i.}}{N}$$

La frecuencia relativa marginal de y_j de Y, viene dada por:

$$f_{.j} = \frac{n_{.j}}{N}$$

- Se cumplen las siguientes relaciones entre las frecuencias de distribución marginales:

1ª) La suma de frecuencias absolutas marginales de la variable X, es igual al número de observaciones que componen la muestra

2ª) La suma de las frecuencias relativas marginales de la variable X, es igual a 1.

3ª) Las dos propiedades anteriores se cumplen también para la variable Y.

c) Distribuciones condicionadas

Consideremos a los $n_{.j}$ individuos de la población que representan la modalidad y_j de la variable Y, y obsérvese la columna j-esima de la tabla. Sus $n_{.j}$ elementos constituyen una población, que es un subconjunto de la población total. Sobre este subconjunto se define la distribución de X condicionada por y_j, que se representa por X / y_j ;su frecuencia absoluta se representa por n_i / j , y su frecuencia relativa por f_i / $_j$, para i = 1, 2, 3, …., r siendo $f_i / _j = \frac{n_{ij}}{n_{.j}}$

El razonamiento es análogo cuando condicionamos la variable Y a un determinado valor de X, es decir Y /x_i

Ejemplo:

Sea X= salario en u.m.

Sea Y = antigüedad en la empresa (años)

X / Y	**1**	**3**	**5**	**7**	**9**	**11**	**ni.**	**fi.**
90	1	2	1	1	0	0	5	0,053
110	2	4	4	5	2	1	18	0,189
130	1	7	3	1	2	0	14	0,147
150	4	6	6	4	3	0	23	0,242
170	2	3	4	6	4	1	20	0,211

190	0	0	2	5	5	3	15	0,158
n.j	10	22	20	22	16	5	95	1
f.j	0,105	0,232	0,211	0,232	0,168	0,053		1

¿ Cuál es la distribución de la retribución, pero únicamente de los empleados con una antigüedad de 5 años?, es decir ¿ cual es la distribución condicionada de la variable X condicionada a que Y sea igual a 5?

X / Y	**ni/ y=5**	**fi/ y=5**
90	1	1/20
110	4	4/20
130	3	3/20
150	6	6/20
170	4	4/20
190	2	2/20
n.j	20	1

- Covarianza

La covarianza mide la forma en que varía conjuntamente dos variables X e Y

En el estudio conjunto de dos variables, lo que nos interesa principalmente es saber si existe algún tipo de relación entre ellas. Veremos ahora una medida descriptiva que sirve para medir o cuantificar esta relación:

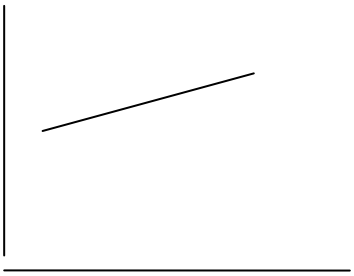
$$S_{xy} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(x_i - \overline{x})(y_j - \overline{y})n_{ij}}{N}$$

Si $S_{XY} > 0$ hay dependencia directa (positiva), es decir las variaciones de las variables tienen el mismo sentido

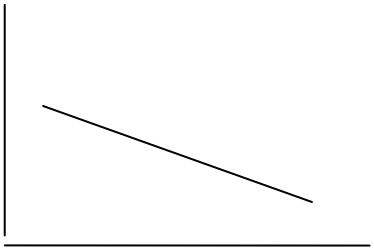
Si $S_{XY} = 0$ las variables están incorreladas, es decir no hay relación lineal, pero podría existir otro tipo de relación.

Si $S_{XY} < 0$ hay dependencia inversa o negativa, es decir las variaciones de las variables tienen sentido opuesto.

Gráficamente, indicaría la Covarianza, que los datos, se ajustan a una recta, en los siguientes casos:



$S_{xy} > 0$



$S_{xy} < 0$

- Otra forma de calcular la Covarianza sería: $S_{xy} = m_{11} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j n_{ij}}{N} - \bar{x}\bar{y}$.
  Será la que utilizaremos en la práctica.
- La covarianza no es un parámetro acotado, y puede tomar cualquier valor real, por lo que su magnitud no es importante; lo significativo es el signo que adopte la misma.

Ejemplo:
Sea X el tiempo de vida de un insecto (años) e Y la longitud del mismo, ¿ podrías deducir si existe relación entre la "edad" del insecto y su tamaño.

X / Y	**2**	**3**	**4**	**ni.**
1	3	1	0	4
2	1	3	1	5
3	0	1	3	4
n.j	4	5	4	13

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_{i.}}{N} = \frac{1*4+2*5+3*4}{13} = 2 \text{ años}$$

$$\bar{y} = \frac{\sum_{j=1}^{s} y_j n_{.j}}{N} = \frac{2*4+3*5+4*4}{13} = 3 \text{ cms}$$

$$S_{xy} = \frac{1*2*3+1*3*1+1*4*0+2*2*1+2*3*3+2*4*1+3*2*0+3*3*1+3*4*3}{13} - 2*3 = 0.461$$

Al tener la covarianza entre ambas variables signo positivo, podemos deducir que existe una relación directa o positiva entre ambas variables, es decir, cuando aumenta la " edad " del insecto también aumenta su tamaño.

3.2.2.Tablas de contingencia

Cuando tenemos la información de 2 variables de tipo cualitativo o de una variable cualitativa y otra cuantitativa, se dispone de una tabla de contingencia. Nos limitaremos al caso de 2 variables. Es una tabla de doble entrada en la que en las filas se ubican las modalidades de una de las variables (atributos) y en las columnas las del otro; en las celdas resultantes del cruce de las filas y las columnas se incluye el número de elementos de la distribución que presentan ambas modalidades.

Si se tiene información de N elementos acerca de las variables A y B de tal forma que presentan " r " y " s " modalidades respectivamente, la tabla de contingencia sería de la forma:

B / A	**B_1**	**B_2**	**.....**	**B_j**	**.....**	**B_s**	**n_i .**	**fi .**
A_1	**n11**	**n12**	**.....**	**n1j**	**.....**	**n1s**	**n_1 .**	**f1 .**
A_2	**n21**	**n22**	**.....**	**n2j**	**.....**	**n2s**	**n_2 .**	**f_2 .**
. . .	**. .**	**. .**	**. .... .**	**. .**	**. .... .**	**. .**	**. .**	**. .**
A_i	**ni1**	**ni2**	**.....**	**nij**	**.....**	**nis**	**n_i .**	**fi .**
. . .	**. . .**	**. . .**	**. .... .**	**. . .**	**. .... .**	**. . .**	**. . .**	**. . .**
A_r	**nr1**	**nr2**	**.....**	**nrj**	**.....**	**nrs**	**n_r .**	**f_r .**
n. s	**n. 1**	**n. 2**	**.....**	**n. j**	**.....**	**n. s**	**N**	
f. s	**f. 1**	**f. 2**	**.....**	**f. j**	**.....**	**f. s**		**1**

tabla de contingencia r x s

n_{ij}= número de elementos de la distribución que presentan la modalidad i – ésima del atributo A y la modalidad j – esima del atributo B.

$n_{i.}$= n_{i1+} n_{i2+} ... + n_{is} --→ número de elementos de la distribución con la i – ésima modalidad del atributo A.

Como a las variables cualitativas no se les puede someter a operaciones de sumas, restas y divisiones, al venir expresadas en escalas nominales u ordinales no tiene sentido hablar de medias marginales, condicionadas, varianzas, etc; si podríamos calcular la moda en el caso de que se empleara una escala nominal y de la mediana si utilizamos escalas ordinales.

3.3. Dependencia e independencia

3.3.1.Independencia

Cuando no se da ningún tipo de relación entre 2 variables o atributos, diremos que son independientes.
Dos variables X e Y, son independientes entre si, cuando una de ellas no influye en la distribución de la otra condicionada por el valor que adopte la primera. Por el contrario existirá dependencia cuando los valores de una distribución condicionan a los de la otra.
Dada dos variables estadísticas X e Y, la condición necesaria y suficiente para que sean independientes es:

$$\frac{n_{ij}}{N} = \frac{n_{i.}}{N}\frac{n_{.j}}{N} \forall i,j$$

Propiedades:

1ª) Si X es independiente de Y, las distribuciones condicionadas de X/Y_j son idénticas a la distribución marginal de X.
2ª) Si X es independiente de Y, Y es independiente de X.
3ª) Si X e Y son 2 variables estadísticamente independientes, su covarianza es cero. La recíproca de esta propiedad no es cierta, es decir, la covarianza de 2 variables puede tomar valor cero, y no ser independientes.

3.3.2.Dependencia funcional (existe una relación matemática exacta entre ambas variables)

El carácter X depende del carácter Y, si a cada modalidad y_j de Y corresponde una única modalidad posible de X. Por lo tanto cualquiera que sea j, la frecuencia absoluta n_{ij} vale cero salvo para un valor de i correspondiente a una columna j tal que n_{ij} = n.j

Cada columna de la tabla de frecuencias tendrá, por consiguiente, un único término distinto de cero. Si a cada modalidad x_i de X corresponde una única

modalidad posible de Y, será Y dependiente de X. La dependencia de X respecto de Y no implica que Y dependa de X.

Para que la dependencia sea recíproca, los caracteres X e Y deben presentar el mismo número de modalidades (debe ser n=m) y en cada fila como en cada columna de la tabla debe haber uno y solo un término diferente de cero.

Sea X el salario de un empleado e Y la antigüedad del mismo en la empresa

X \ Y	**1**	**3**	**5**	**7**	**9**
100	15	0	0	0	0
120	0	20	0	0	0
140	0	0	30	0	0
160	0	0	0	25	0
180	0	0	0	0	10

Dependencia funcional recíproca: X depende de Y e Y depende de X

X \ Y	**1**	**3**	**5**	**7**	**9**	**10**
100	15	0	0	0	0	0
120	0	20	0	0	0	0
140	0	0	30	0	12	0
160	0	0	0	25	0	0
180	0	0	0	0	0	9

Y depende de X pero X no depende de Y

3.3.3.Dependencia estadística (existe una relación aproximada)

Existen caracteres que ni son independientes, ni se da entre ellos una relación de dependencia funcional, pero si se percibe una cierta relación de dependencia entre ambos; se trata de una dependencia estadística.
Cuando los caracteres son de tipo cuantitativo, el estudio de la dependencia estadística se conoce como el problema de " regresión ", y el análisis del grado de dependencia que existe entre las variables se conoce como el problema de correlación.

3.4.Regresión y correlación lineal simple

3.4.1.Introducción a la regresión lineal simple

Cuando se estudian dos características simultáneamente sobre una muestra, se puede considerar que una de ellas influye sobre la otra de alguna manera. El objetivo principal de la regresión es descubrir el modo en que se relacionan.

Por ejemplo, en una tabla de pesos y alturas de 10 personas

Altura	175	180	162	157	180	173	171	168	165	165
Peso	80	82	57	63	78	65	66	67	62	58

se puede suponer que la variable "Altura" influye sobre la variable "Peso" en el sentido de que pesos grandes vienen explicados por valores grandes de altura (en general).

De las dos variables a estudiar, que vamos a denotar con X e Y, vamos a llamar a la X VARIABLE INDEPENDIENTE o EXPLICATIVA, y a la otra, Y, le llamaremos VARIABLE DEPENDIENTE o EXPLICADA.

En la mayoría de los casos la relación entre las variables es mutua, y es difícil saber qué variable influye sobre la otra. En el ejemplo anterior, a una persona que mide menos le supondremos menor altura y a una persona de poca altura le supondremos un peso más bajo. Es decir, se puede admitir que cada variable influye sobre la otra de forma natural y por igual. Un ejemplo más claro donde distinguir entre variable explicativa y explicada es aquel donde se anota, de cada alumno de una clase, su tiempo de estudio (en horas) y su nota de examen. En este caso un pequeño tiempo de estudio tenderá a obtener una nota más baja, y una nota buena nos indicará que tal vez el alumno ha estudiado mucho. Sin embargo, a la hora de determinar qué variable explica a la otra, está claro que el "tiempo de estudio" explica la "nota de examen" y no al contrario, pues el alumno primero estudia un tiempo que puede decidir libremente, y luego obtiene una nota que ya no decide arbitrariamente. Por tanto,

X = Tiempo de estudio (variable explicativa o independiente)
Y = Nota de examen (variable explicada o dependiente)

El problema de encontrar una relación funcional entre dos variables es muy complejo, ya que existen infinidad de funciones de formas distintas. El caso más sencillo de relación entre dos variables es la relación LINEAL, es decir que

$$\boxed{Y = a + b\,X}$$

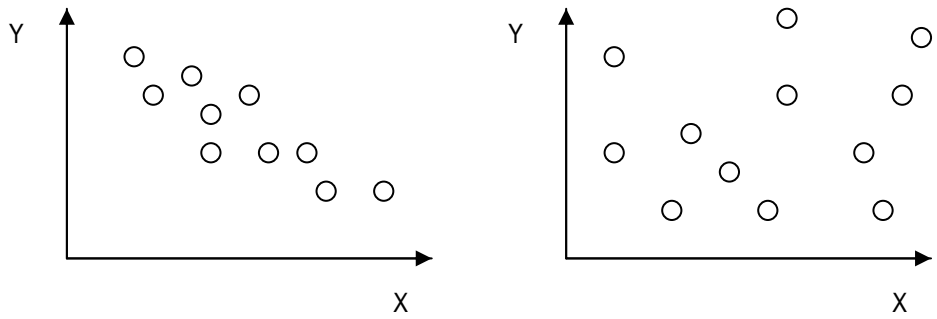
(es la ecuación de una recta) donde a y b son números, que es el caso al que nos vamos a limitar.

Cualquier ejemplo de distribución bidimensional nos muestra que la relación entre variables NO es EXACTA (basta con que un dato de las X

tenga dos datos distintos de Y asociados, como en el ejemplo de las Alturas y Pesos, que a 180 cm. de altura le correspondía un individuo de 82 kg. y otro de 78 kg.).

- Diagrama de dispersión o nube de puntos

En un problema de este tipo, se observan los valores (x_i,y_j) y se representan en un sistema de ejes coordenados, obteniendo un conjunto de puntos sobre el plano, llamado " diagrama de dispersión o nube de puntos ".

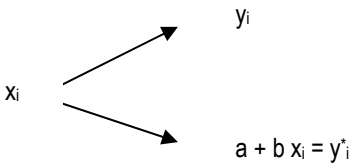


En los diagramas de arriba se puede observar cómo en el de la izquierda, una línea recta inclinada puede aproximarse a casi todos los puntos, mientras que en el otro, cualquier recta deja a muchos puntos alejados de ella. Así pues, el hacer un análisis de regresión lineal sólo estaría justificado en el ejemplo de la izquierda.

Como se puede ver en ambos diagramas, ninguna recta es capaz de pasar por todos los puntos, y seguir siendo recta. De todas las rectas posibles, la RECTA DE REGRESIÓN DE Y SOBRE X es aquella que minimiza un cierto error, considerando a X como variable explicativa o independiente y a Y como la explicada o dependiente.

- Recta de mínimos cuadrados o recta de regresión de Y sobre X ($y^* = a + b\,x$)

Sea $y = a + b\,x$ una recta arbitraria. Para cada dato de X, es decir, para cada x_i de la tabla tenemos emparejado un dato de Y llamada y_i, pero también tenemos el valor de sustituir la x_i en la ecuación de la recta, al que llamaremos y^*_i.



Cuando se toma el dato x_i, el error que vamos a considerar es el que se comete al elegir y^*_i en lugar del verdadero y_i .Se denota con e_i y vale

$e_i = y_i - y^*_i$

Esos errores pueden ser positivos o negativos, y lo que se hace es escoger la recta que minimice la suma de los cuadrados de todos esos errores, que es la misma que la que minimiza la varianza de los errores.

Usando técnicas de derivación se llega a que, de todas las rectas y = a + b x, con a y b números arbitrarios, aquella que minimiza el error elegido es aquella que cumple

$$a = \bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} \qquad \text{y} \qquad b = \frac{s_{xy}}{s_x^2} \qquad \text{por lo tanto} \quad a = \bar{y} - b\bar{x}$$

Así pues, sustituyendo en y = a + b x, la ecuación de la recta de regresión de Y sobre X es

$$y* = \left(\bar{y} - \frac{s_{xy}}{s_x^2} \cdot \bar{x} \right) + \left(\frac{s_{xy}}{s_x^2} \right) \cdot x \qquad \text{es decir} \qquad y^* = a + b\bar{x}$$

y recolocando los términos se puede escribir de la forma

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

- Recta de regresión de X sobre Y

Si se hubiese tomado Y como variable independiente o explicativa, y X como dependiente o explicada, la recta de regresión que se necesita es la que minimiza errores de la X. Se llama RECTA DE REGRESIÓN DE X SOBRE Y y se calcula fácilmente permutando los puestos de x e y, obteniéndose

$$x - \bar{x} = \frac{s_{xy}}{s_y^2} \cdot (y - \bar{y}) \quad \textbf{es decir} \qquad x^* = a' + b' y$$

Sabiendo que : $b' = \dfrac{S_{xy}}{S_y^2}$ y que $a' = \bar{x} - b'\bar{y}$

PROPIEDADES:

- Ambas rectas de regresión pasan por el punto ($\overline{x}, \overline{y}$)
- La pendiente de la recta de regresión de Y sobre X es " b " y la de X sobre Y es " b ´ ". Dado que las varianzas son positivas por definición, el signo de las pendientes será el mismo que el de la covarianza, y así, las rectas serán ambas crecientes o decrecientes, dependiendo de si la covarianza es positiva o negativa, respectivamente, es decir b y b´ tendrán el mismo signo.
- Los términos de las rectas a y a´ constituyen los orígenes de las rectas, es decir, son los valores que adoptan respectivamente y* ó x* cuando x o y toman el valor cero en sus correspondientes rectas de regresión.
- Las rectas de regresión las emplearemos para realizar predicciones acerca de los valores que adoptaran las variables.
- Puede darse el caso, de no existencia de correlación lineal entre las variables, lo cual no implica que no existan otro tipo de relaciones entre las variables estudiadas: relación exponecial, relación parabólica, etc.

3.4.2.Correlación lineal simple (r ó R)

Para ver si existe relación lineal entre dos variables X e Y, emplearemos un parámetro que nos mida la fuerza de asociación lineal entre ambas variables. La medida de asociación lineal mas frecuentemente utilizada entre dos variables es " r " o coeficiente de correlación lineal de Pearson; este parámetro se mide en términos de covarianza de X e Y.

$$R = \frac{S_{xy}}{S_x S_y} \qquad -1 \le R \le 1$$

- Si R = 1, existe una correlación positiva perfecta entre X e Y
- Si R = -1, existe una correlación negativa perfecta entre X e Y
- Si R = 0, no existe correlación lineal, pudiendo existir otro tipo de relación
- Si $-1 \prec R \prec 0$, existe correlación negativa y dependencia inversa, mayor cuanto más se aproxime a - 1.
- Si $0 \prec R \prec 1$, existe correlación positiva, y dependencia directa, mayor cuanto más se aproxime a 1.

- **<u>Varianza residual y varianza explicada por la regresión. Coeficiente de determinación lineal (R^2)</u>**

Si tenemos dos variables X e Y relacionadas linealmente, parte de la variabilidad de la variable Y, vendrá explicada por variaciones de X (variabilidad explicada por el modelo) , mientras que el resto responderá a variaciones de fenómenos relacionados con la variable Y o con el azar (variabilidad no explicada por el modelo) .

Por tanto nos conviene disponer de una medida que indique el porcentaje de la variabilidad de la variable explicada que se debe a la variabilidad de la variable explicativa. Esta medida es el coeficiente de determinación lineal (R^2) , y si su valor es alto nos indicará que el ajuste lineal efectuado es bueno.

En la regresión lineal de Y sobre X, la varianza de la variable Y, puede descomponerse en la suma de 2 varianzas:

$$S_y^2 = S_r^2 + S_e^2$$

donde:

S_y^2 es la varianza total de la variable Y

S_r^2 es la varianza explicada o variabilidad de Y explicada por la **r**egresión. $\boxed{S_r^2 = bS_{xy}}$

S_e^2 es la varianza residual (**e**) o variabilidad de Y no explicada por la **r**egresión. $\boxed{S_e^2 = S_y^2 - bS_{xy}}$

$$\boxed{R^2 = \frac{S_r^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2}} \qquad 0 \le R^2 \le 1$$

$\boxed{R^2 = \dfrac{S_{xy}^2}{S_x^2 S_y^2}}$ también podemos afirmar que $\boxed{R^2 = b \cdot b'}$

Es una medida de la bondad del ajuste lineal efectuado. Si lo expresamos en porcentaje, dicho coeficiente nos indica el % de la varianza de la variable explicada (Y) que se ha conseguido explicar mediante la regresión lineal.

Si $R^2 = 1$, existe dependencia funcional; la totalidad de la variabilidad de Y es explicada por la regresión.
Si $R^2 = 0$, dependencia nula; la variable explicativa no aporta información válida para la estimación de la variable explicada.
Si $R^2 \geq 0.75$, se acepta el modelo ajustado

- Relación existente entre los coeficientes de determinación y correlación lineal:

$$\boxed{R = \pm\sqrt{R^2}}$$

El signo del coeficiente de correlación lineal será el mismo que el de la covarianza.

3.5. **Estudio de la asociación entre variables cualitativas**.

En el estudio visto de regresión y correlación se ha tratado solo el caso de variables cuantitativas (ingresos, salarios, precios, etc) Con variables de tipo cualitativo se puede construir tablas de contingencia, a través de las

cuales se puede estudiar la independencia estadística entre los distintos atributos.

Si dos atributos son dependientes, se pueden construir una serie de coeficientes que nos midan el grado asociación o dependencia entre los mismos.

Partimos de la tabla de contingencia en la que existen r modalidades del atributo A y s del atributo B. El total de observaciones será:

$$N = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$$

La independencia estadística se dará entre los atributos si : $\frac{n_{ij}}{N} = \frac{n_{i.}}{N} \cdot \frac{n_{.j}}{N}$ para todo i , j ; si esta expresión no se cumple, se dirá que existe un grado de asociación o dependencia entre los atributos.

$$\frac{n_{ij}}{N} \neq \frac{n_{i.}}{N} \cdot \frac{n_{.j}}{N} \quad \text{----} \rightarrow n'_{ij} = \frac{n_{i.} \cdot n_{.j}}{N}$$

El valor n'_{ij} es la frecuencia absoluta conjunta teórica que existiría si los 2 atributos fuesen independientes.
El valor n_{ij} es la frecuencia absoluta conjunta observada.
El coeficiente de asociación o contingencia es el llamado Cuadrado de Contingencia, que es un indicador del grado de asociación:

$$\chi^2 = \sum_i \sum_j \frac{(n'_{ij} - n_{ij})^2}{n'_{ij}} \text{ siendo } n'_{ij} = \frac{n_{i.} \cdot n_{.j}}{N}$$

El campo de variación va desde cero (cuando existe independencia y $n'_{ij} = n_{ij}$), hasta determinados valores positivos, que dependerá de las magnitudes de las frecuencias absolutas que lo componen.

Este inconveniente de los límites variables se eliminará con el empleo del Coeficiente de contingencia de Pearson:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

Varía entre cero y uno. El valor cero se dará en el caso de independencia ($n'_{ij} = n_{ij}$). Cuanto más se aproxime a 1 más fuerte será el grado de asociación entre los dos atributos.

- Estudio de la asociación entre dos atributos

- Para tablas de contingencia 2 x 2

Sean A y B dos variables cualitativas o atributos tales que presentan 2 modalidades cada una. La tabla de contingencia correspondiente es la siguiente:

A \ B	B_1	B_2	
A_1	n_{11}	n_{12}	$n_{1.}$
A_2	n_{21}	n_{22}	$n_{2.}$
	$n_{,1}$	$n_{,2}$	N

A y B son independientes si: $n_{11} = \frac{n_{1.} n_{.1}}{N}$

A las expresiones $\frac{n_{i.} n_{.j}}{N}$ se les denomina frecuencias esperadas y se denotan por n'_{ij} o por E_{ij}

Si finalmente podemos concluir que los dos atributos están asociados, se pueden plantear dos preguntas:
1ª) ¿ Cual es la intensidad de la asociación entre los dos atributos ?
2ª) ¿ Cual es la dirección de la asociación detectada ?

- Asociación perfecta entre dos atributos

Ocurre cuando, al menos, una de las modalidades de uno de los atributos queda determinada por una de las modalidades del otro atributo. Esto ocurre cuando existe algún cero en la tabla 2 x 2. La asociación perfecta puede ser:
a) Asociación perfecta y estricta.
Ocurre cuando dada modalidad de uno de los atributos queda inmediatamente determinada la modalidad del otro. Es decir, cuando $n_{11} = n_{22} = 0$ ó $n_{12} = n_{21} = 0$

Ejemplo:

A= tipo de trabajo (temporal ó indefinido)
B= Sexo (hombre ó mujer)

Sexo \ Tipo trabajo	Temporal	Indefinido
Hombre	20	0
Mujer	0	80

Con estos datos sabemos que si un individuo es hombre el tipo de trabajo sera temporal y si es mujer su contrato será indefinido.

- Asociación perfecta e implicita de tipo 2

Ocurre cuando:

1º) Si se toma la modalidad de un atributo queda determinada la modalidad del otro atributo al que pertenece la observación.

2º) Si se toma la otra modalidad, no queda determinada la modalidad del otro atributo al que pertenece la observación.

Es decir, esta asociación se produce cuando alguna de las frecuencias observada es cero.

Ejemplo:

A= tipo de trabajo (temporal ó indefinido)

B= Sexo (hombre ó mujer)

Sexo \ Tipo trabajo	Temporal	Indefinido
Hombre	5	15
Mujer	0	80

- Si la persona observada es mujer sabremos que su contrato es indefinido; si es varón puede ser indefinido o temporal.
- Si el contrato analizado es temporal pertenecerá a un hombre; si es un contrato indefinido, podrá ser de un hombre o una mujer.

- También podemos delimitar si la asociación es positiva o negativa:

- Asociación positiva

Cuando se verifica que :

a) La modalidad 1 del atributo A está asociada a la modalidad 1 del atributo B
b) La modalidad 2 del atributo A está asociada a la modalidad 2 del atributo B.

- Asociación negativa:

Cuando se verifica que:

a) La modalidad 1 del atributo A está asociada a la modalidad 2 del atributo B
b) La modalidad 2 del atributo A esta asociada a la modalidad 1 del atributo A.

Para medir el sentido de la asociación entre dos atributos emplearemos el indicador Q de Yule:

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

$$\boxed{-1 \le Q \le 1}$$

Si Q = 0, entonces existe independencia

Si Q > 0, entonces existe asociación positiva

Si Q < 0, entonces existe asociación negativa

- Tablas de contingencia R x S

Para determinar la intensidad de dicha asociación, calculamos la V de Cramer, que se define como:

$$V = \sqrt{\frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}}{Nm}}$$

m = min (r-1 , s-1)

$V \in (0,1)$

$$E_{ij} = \frac{n_{i.}n_{.j}}{N}$$

Existirá una mayor intensidad en la asociación entre 2 variables a medida que el indicador adopte valores próximos a 1.