BLOQUE 2 – TEMA 7
RELACIÓN ENTRE VARIABLES : CORRELACIÓN Y REGRESIÓN

En los procesos de investigación en Educación, en muchas ocasiones, nos interesa conocer la posible relación que puede manifestarse entre dos ó más variables.

El concepto de correlación

Kerlinger (1985) afirma que **la esencia de la ciencia son las relaciones entre variables**, que pueden ponerse de manifiesto entre grupos, clases o conjuntos de objetos, pero que no cabe hablar de relaciones entre variables midiendo solamente a un individuo.

Cuando conocemos la relación entre variables, pueden llegar a **formularse predicciones** de los valores de una a partir de la otra.

Las **situaciones que pueden analizarse** son varias:

- estudiar la relación entre dos ó más variables dentro de un mismo grupo de sujetos.
- con dos ó más grupos comprobar la relación entre ellos en una sola variable.
- estudiar una misma variable medida en dos momentos diferentes en una misma muestra.

Una relación simple entre dos series de datos se denomina **correlación** entre dos variables. Nos indica la tendencia entre dos ó más conjuntos de datos a variar de forma conjunta. Tenemos varias posibilidades:

a) **relación perfecta positiva**
cuando al aumentar los valores de una de las variables, los valores de la otra lo hacen en la misma proporción. Ver fig. 7.1 pág. 131. La correlación se expresa como +1

b) **relación imperfecta positiva**
se la conoce como **relación directa** de variables. A valores elevados de una variable le corresponden valores también altos de la otra; y a la inversa, los que puntúan bajo coinciden en las dos variables. Ver fig. 7.2 pág. 131. La correlación se sitúa entre los valores 0 y +1

c) **relación perfecta negativa**
se da una relación inversa entre las variables, de tal forma que al aumentar los valores de una de ellas, los de la otra disminuyen y lo hacen en la misma proporción. Ver fig. 7.3 pág. 132. La correlación se expresa como -1

d) **relación imperfecta negativa**
llamada **relación inversa** entre variables, lo que supone que las puntuaciones altas en una variable se corresponden con las bajas en la otra. Ver fig. 7.4 pág. 132. La correlación se sitúa entre los valores 0 y -1

e) **relación nula o ausencia de relación**
se da cuando dos variables son independientes una de la otra. Puede afirmarse que las puntuaciones de las dos variables se deben a factores aleatorios. La correlación se expresa por 0.

El coeficiente de correlación simple y su interpretación

El valor del coeficiente nos marca el valor de la **covariación ó variación conjunta** de dos series de datos. Puede indicar una **relación directa** entre variables (valores positivos) ó **inversa** (valores negativos), por loq eu su expresión se encuentra entre -1 y +1.

Cómo se interpreta un coeficiente de correlación?. Teniendo en cuenta tres aspectos:

- *el tipo de variables que se relacionan*: entre variables del mismo tipo, mayor correlación.
- *la variabilidad del grupo*: a mayor variabilidad entre los grupos y dentro de ellos, mayor será la correlación. ante un coeficiente de 0,70, el obtenido en el grupo más homogéneo (menor variabilidad) se identifica con una mayor intensidad de la correlación.
- *la finalidad a la que se destina el coeficiente*: si valoramos la fiabilidad de un instrumento de medida, las correlaciones deben superar el 0,85; mientras que un coeficiente de 0,60 es suficiente si valoramos la validez del instrumento.

Se aceptan las interpretaciones de la tabla 7.1 pág. 134.

También podemos interpretarlo mediante el **coeficiente de determinación (d)**, que se interpreta como el porcentaje de varianza de una variable explicada por la otra:

$$d = (r_{xy}{}^2)100$$

Si r_{xy} = 0,70 → d = $(0,70)^2$. 100 = 49 → 49% de varianza explicada.

La elección de los diferentes coeficientes de correlación depende de dos aspectos: *el nivel de medida de las variables y la categoría de las mismas*. Así, tenemos:

INDICE	NIVEL DE MEDIDA	CATEGORÍAS DE LAS VARIABLES
Pearson → r_{xy}	Intervalo	Ambas continuas y normales
Coeficiente de contingencia → C	Nominal	Atributos, en dos ó más categorías
Spearman → r_s	Ordinal	Continuas, por rangos
Biserial → r_b	Intervalo	Una continua y la otra dicotomizada
Biserial puntual → r_{bp}	Intervalo	Continua y dicotómica
Tetracórica → r_t	Intervalo	Ambas continuas y dicotomizadas
Phi → ϕ	Nominal	Ambas dicotómicas ó por atributos (dos únicas categorías)

El coeficiente de correlación de Pearson (r_{xy})

Se utiliza cuando las dos variables que se relacionan son cuantitativas, medidas a nivel de intervalo, se distribuyen normalmente y están linealmente relacionadas.

típicas →
$$r_{xy} = \frac{\sum z_x . z_y}{N}$$

diferenciales →
$$r_{xy} = \frac{\sum x.y}{N.S_x.S_y} = \frac{\sum x.y}{\sqrt{\sum x^2}.\sqrt{y^2}}$$

directas →
$$r_{xy} = \frac{N.\sum XY - \sum X \sum Y}{\sqrt{N.\sum X^2 - (\sum X)^2}.\sqrt{N.\sum Y^2 - (\sum Y)^2}}$$

directas →
$$r_{xy} = \frac{\sum XY - N\bar{x}\bar{y}}{\sqrt{\sum X^2 - N(\bar{x})^2}.\sqrt{\sum Y^2 - N(\bar{y})^2}}$$

Para el cálculo de estas fórmulas se construye la siguiente tabla:

Sujetos	Var X	Var Y	X²	Y²	XY		x	y	x^2	y^2	xy
	$\sum$	$\sum$	$\sum$	$\sum$	$\sum$				$\sum$	$\sum$	$\sum$

Ver ejemplo en tabla 7.2 pág. 136

El coeficiente de correlación de Spearman (r_s)

En muchas de las variables que utilizamos frecuentemente en el campo educativo no es posible alcanzar unos niveles de medida muy precisos. Es entonces cuando hemos de recurrir a emplear los puestos que ocupan las puntuaciones en una serie ordenada.

En medidas ordinales hemos de recurrir a los rangos.

Para hacer la transformación de las puntuaciones directas en rangos se siguen estos pasos:

- se asigna el rango 1 a la posición más alta, rango 2 a la siguiente en descenso, 3 a la siguiente, y así hasta que el último rango asignado coincida con N.
- cuando tengamos más de una puntuación similar, el rango se calcula mediante la $\bar{x}$ de las posiciones que corresponderían a esos sujetos.
- el criterio de asignación de rangos empleado en una de las variables, debe ser el mismo que en la otra, pues se trata de pares de puntuaciones que van asociadas y no son independientes.

$$r_s = 1 - \frac{6 \sum D^2}{n^3 - n}$$

n → número de sujetos.
D → $[R(X) - R(Y)]$ → diferencia de rangos ó posiciones que ocupa un mismo sujeto en dos variables distintas.

Ver tablas 7.3 y 7.4 págs. 138 y 139
Para el cálculo se construye la siguiente tabla:

Sujetos	Var X	Var Y	R(X)	R(Y)	D	D²
	$\sum$	$\sum$				$\sum$

<u>Coeficiwente de contingencia (C)</u>
En el caso de **variables de atributo ó nominales**, es preferible utilizar la expresión **grado de asociación** en vez de grado de correlación. Se utiliza en aquellos casos en que se recogen datos de variables clasificadas en categorías, como ocurre con las tablas de contingencia, en las que se asignan sujetos a grupos y categorías en cada una de las variables.
La fórmula es : $C = \sqrt{\frac{\chi^2}{n + \chi^2}}$ en la cual:

$$\chi^2 = \sum_1^f \sum_1^c \frac{(fo - fe)^2}{fe}$$

fo → frecuencia observada que aparece en cada celdilla de la tabla de contingencia.
fe → frecuencia esperada teórica que refleja el grado de asociación.
Las fe pueden calcularse mediante la fórmula siguiente:

$$\text{fe} = \frac{\sum f_{filas} \cdot \sum f_{columnas}}{\sum f_{total}}$$

Debemos comenzar por calcular **el valor de fe** en cada una de las celdillas de la tabla de contingencia.
Ver fig. 7.5 y 7.6 pág. 141 y 142.
Seguidamente procedemos a **calcular χ²,** que representa el grado de discrepancia que se manifiesta entre las frecuencias observadas ó empíricas (**fo**) y las frecuencias esperadas ó aleatorias (**fe**).
Para finalizar, **se calcula el valor de C**. Ver ejemplo en págs.. 141 y 142.
El coeficiente nunca alcanzará el valor de 1, y para su interpretación se recurre al **valor de Cmáx**., que solo es válido su cálculo cuando las tablas de contingencia sean cuadradas, es decir, el mismo número de filas que de columnas.
Como comprobación, la ∑fo y la ∑fe dentro de la misma fila ó columna, deben ser iguales.

<u>El coeficiente de correlación biserial puntual (rbp)</u>
Cuando buscamos el grado de relación entre una **variable cuantitativa y otra auténticamente dicotómica**, debemos recurrir al r_{bp}. En realidad, es una extensión del coeficiente de correlación de Pearson.

$$r_{bp} = \frac{|\bar{x}_p - \bar{x}_t|}{s_t} \cdot \sqrt{\frac{p}{q}} \qquad r_{bp} = \frac{|\bar{x}_p - \bar{x}_q|}{s_t} \cdot \sqrt{pq}$$

El numerador de ambas fórmulas se toma en valores absolutos con el fin de evitar los valores negativos para la correlación. Generalmente, los datos de la variable continua suelen agruparse en **intervalos de clase**, tal y como aparece en la tabla 7.7 pág. 144.
Hemos de realizar unos **cálculos previos**:

- $\bar{x}$ de cada uno de los grupos y $\bar{x}$ del total de sujetos.
- St del conjunto de puntuaciones.
- proporción de cada uno de los grupos en relación al total.
- finalmente, completar la siguiente tabla:

Intervalo	fp	fq	ft	Xc	Xcfp	Xcfq	Xcft	Xc^2ft
	$\sum$	$\sum$	$\sum$		$\sum$	$\sum$	$\sum$	$\sum$

$$\bar{X} = \frac{\sum fiXi}{N} \qquad S^2 = \frac{\sum fiXi^2}{N} - (\bar{X})^2$$

Xc $\rightarrow$ marca de clase del intervalo
p y q $\rightarrow$ categorías de la variable dicotómica

La r_{bp} se utiliza en el análisis de los elementos de pruebas objetivas, especialmente en aquellos casos en que la respuesta no admite nada más que los valores de acierto ó error. En estos casos, el r_{bp} es un **índice de la homogeneidad** de tal elemento ó ítem con la puntuación global de la prueba.

<u>El corficiente phi (ϕ)</u>
Se emplea cuando buscamos la existencia de relaciones entre **dos variables dicotómicas**, aunque de forma excepcional puede utilizarse en variables dicotomizadas.
Una variable es dicotomizada cuando transformamos la variable continua a una escala de dos categorías, siendo muy frecuentemente el punto de dicotomización, la **Md**. Ver tabla 7.9 pág. 145

$$\phi = \frac{|BC-AD|}{\sqrt{(A+B).(A+C).(C+D).(B+D)}}$$

		X: p	X: q	
Y	p	A	B	A+B
	q	C	D	C+D
		A+C	B+D	∑total

Es la aplicación a variables dicotómicas del coeficiente r_{xy} de Pearson.
Su valor será 1 (correlación perfecta) cuando las frecuencias de las casillas de las diagonales positiva ó negativa sean nulas. en los demás casos se estudia su valor comparándolo con el **valor del ϕmáx**., muy importante como punto de referencia y de interpretación:
donde:

- $f_m \rightarrow$ frecuencia marginal mayor de las cuatro
- $f_{1-m} \rightarrow$ frecuencia marginal complementaria de esa misma variable
- $f_p \rightarrow$ frecuencia marginal mayor de la otra variable
- $f_{1-p} \rightarrow$ frecuencia marginal complementaria de la anterior

$$\phi_{máx} = \sqrt{\left(\frac{f_{1-m}}{f_m}\right).\left(\frac{f_p}{f_{1-p}}\right)}$$

El valor hallado de ϕ no se puede interpretar como r_{xy} , sino que hemos de referirlo al valor máximo alcanzable con las frecuencias marginales que manejemos. Un ejemplo:

		X: p	X: q	
Y	p	A 21	B 42	63
	q	C 30	D 5	35
		51	47	98

$$\phi = \frac{|30\ .42-21\ .5|}{\sqrt{63\ .35\ .51\ .47}} = \frac{1155}{2299} = 0{,}502$$

$$\phi_{máx} = \sqrt{\left(\frac{35}{63}\right).\left(\frac{51}{47}\right)} = \sqrt{0{,}555\ .1{,}085} = 0{,}776$$

También puede calcularse si disponemos de la distribución de frecuencias de las variables.
Veamos un ejemplo:

Sujeto	X Sexo H:1 // M:0	Y Selectividad Apto:1 // No Apto:0
1	0	0
2	1	1
3	0	1
4	0	0
5	1	1
6	1	0
7	0	0
8	1	1
9	0	0
10	0	1
11	0	0
12	1	1
	$\sum$=5	$\sum$=6

$$\phi = \frac{p_{xy} - p_x.p_y}{\sqrt{p_x.q_x.p_y.q_y}}$$

$p_x \rightarrow$ proporción de 1 en X $\rightarrow$ 5/12 = 0,4167
$q_x \rightarrow$ proporción de 0 en Y $\rightarrow$ 7/12 = 0,5833
$p_y \rightarrow$ proporción de 1 en Y $\rightarrow$ 6/12 = 0,50
$q_y \rightarrow$ proporción de 0 en Y $\rightarrow$ 1- p_y = 0,50
$p_{xy} \rightarrow$ proporción de 1 tanto en X como en Y $\rightarrow$ 4/12 = 0,33

$$\phi = \frac{0{,}33 - (0{,}4167)\,.0{,}50}{\sqrt{(0{,}4167).(0{,}5833).(0{,}50).(0{,}50)}} = 0{,}507$$

Agrupando datos, tendríamos que:

		X p	X q	
Y	p	A 4	B 2	6
Y	q	C 1	D 5	6
		5	7	12

$$\phi = \frac{|2-20|}{\sqrt{5\,.7\,.6\,.6}} = \frac{18}{6\sqrt{35}} = \frac{3}{\sqrt{35}} = 0{,}507$$

$$\phi_{max} = \sqrt{\frac{5}{7}.\frac{6}{6}} = \sqrt{0{,}7142} = 0{,}845$$

<u>Coeficiente de correlación tetracórico (r_t)</u>

Se emplea en aquellos casos en que las **dos variables** son de tipo cuantitativo y continuo, pero **nos interesa dicotomizarlas**, dividiendo las puntuaciones de cada variable en dos categorías y tomando como **criterio de categorización** generalmente la **Md**.

Podemos entonces formalizar una **tabla de contingencia de 2x2**:

		Y -	Y +
X	-	A	B
X	+	C	D

A $\rightarrow$ + en X; - en Y
B $\rightarrow$ + en X; + en Y
C $\rightarrow$ - en X; - en Y
D $\rightarrow$ - en X; + en Y

$$r_t = \frac{BC}{AD}$$

En el **numerador** siempre el producto de la diagonal de igual signo, y en el **denominador** siempre el producto de la diagonal de signos distintos.
En aquellos casos que BC < AD, buscaremos en las tablas el valor de (AD/BC) y al coeficiente resultante le pondremos un signo menos.
También hay un **procedimiento numérico** para hallar el valor de este coeficiente. Viene dado por la fórmula:

$$r_t = cos\left(\frac{180\,.\sqrt{AD}}{\sqrt{BC} + \sqrt{AD}}\right)$$

Ver ejemplo resuelto en págs.. 146 y 147.

Coeficiente de relación biserial (r_b)

Se utiliza cuando se trata de establecer la relación que existe entre **una variable cuantitativa y otra dicotomizada** (una variable continua que ha sido dividida en dos categorías de forma artificial). La variable continua se presenta en **escala de intervalo**.

$$r_b = \frac{|\bar{x}_p - \bar{x}_t|}{s_t} \cdot \frac{p}{Y} \qquad r_b = \frac{|\bar{x}_p - \bar{x}_q|}{s_t} \cdot \frac{p \,.\, q}{Y}$$

siendo **Y la ordenada en el punto z** en que se establece el cambio de categoría y que se obtiene de las tablas con el valor de la proporción de p $\rightarrow (n_p/n_t)100$

Hemos de realizar unos cálculos previos:

- $\bar{x}$ de cada uno de los grupos y $\bar{x}$ del total de sujetos
- S_t del total de puntuaciones
- proporciones de las dos categorías de dicotomización (**p** y **q**)
- valor de Y en tablas de la curva normal

Para efectuar los cálculos, completamos la siguiente tabla:

Intervalo	fp	fq	ft	Xc	Xcfp	Xcfq	Xcft	Xc^2ft
	$\sum$	$\sum$	$\sum$		$\sum$	$\sum$	$\sum$	$\sum$

$$\bar{X} = \frac{\sum fiXi}{N} \qquad s^2 = \frac{\sum fiXi^2}{N} - (\bar{x})^2$$

Xc $\rightarrow$ marca de clase del intervalo

p y q $\rightarrow$ categorías de la variable dicotómica

Ver ejemplo 7.1 en pág 148

También puede calcularse si disponemos de la distribución de frecuencias sin agrupar en intervalos:

Sujeto	X Tiempo de estudio (en min.)	X^2	Y Puntuaciones: suficiente = 1 Insuficiente=0
1	20	400	1
2	25	625	0
3	30	900	1
4	16	256	0
5	10	100	0
6	40	1600	1
7	18	324	0
8	50	2500	1
9	43	1849	1
10	49	2401	1
	$\sum$	$\sum$	

$$r_b = \frac{\bar{x}_1 - \bar{x}_0}{s_x} \cdot \frac{n_1 \,.\, n_0}{Y.n.\sqrt{n^2 - n}}$$

$\bar{x}_1$ y $\bar{x}_2$ $\rightarrow$ medias de X a las que se les asignó un 1 y un 0 respectivamente.

Sx $\rightarrow$ desviación típica de las puntuaciones.

$$s_x^2 = \frac{\sum X^2}{N} - (\bar{x})^2$$

n_1 y n_0 $\rightarrow$ número de 1 y de 0 en Y. Nótese que ($n_1 + n_0 = n$)

Y $\rightarrow$ ordenada de la distribución normal, en el punto donde se encuentra el porcentaje $(n_1/n)100$ del área bajo la curva.

Cálculos previos:

$n_1 = 6$ $\quad \bar{X}_1 = \frac{\sum X_1}{6} = \frac{232}{6} = 38{,}66$ $\quad s_x = \sqrt{\frac{10955}{10} - (30{,}10)^2} = \sqrt{189{,}49} = 13.76$

$n_o = 4$ $\quad \bar{X}_0 = \frac{\sum X_0}{4} = \frac{69}{4} = 17{,}25$

$n = 10$ $\quad \bar{X} = \frac{\sum X_i}{10} = \frac{301}{10} = 30{,}10$

Para $\left(\frac{n_1}{n}\right)100 \rightarrow$ 60% de **p**, (**q** = 1- p = 40%), en tablas pág. 336, Y = 0,3867 (por aproximación)

$$r_b = \frac{38{,}66 - 17{,}25}{13{,}76} \cdot \frac{6 \,.\, 4}{(0{,}3867) \,.10 \,.\sqrt{1000 - 10}} = 1{,}555 \,.0{,}654 = 1{,}02$$

El coeficiente r_b puede ser menor que -1 y mayor que +1, en cuyo caso las puntuaciones X no se distribuyen normalmente o son fluctuaciones de muestreo cuando **n** es pequeño, como es nuestro caso del ejemplo.

La regresión lineal simple

La interpretación de los coeficientes de correlación se basa en la intensidad ó el grado de esa relación, desde valores próximos a 0 hasta los cercanos a 1, que indica la mayor intensidad.

Estos valores permiten conocer la *varianza compartida*, es decir, la parte de la variabilidad de una de ellas explicada por la otra. Para su cálculo se eleva el valor del coeficiente al cuadrado y se multiplica por 100, obteniéndose el llamado *coeficiente de determinación*.

Mediante este coeficiente de determinación podemos estimar los valores de una variable conociendo los valores de la otra. es lo que conocemos como regresión lineal simple, cuya función principal nos permite la predicción.

Lo que pretendemos es predecir puntuaciones en una variable sin aplicar el instrumento de medida, simplemente conociendo la relación de esa variable con otra y de la que tenemos los resultados de determinados sujetos.

Obtenemos así una relación positiva y prácticamente lineal, conocida como la **línea de regresión** y que se emplea para llevar a cabo la predicción ó estimación de los valores de **una variable Y´ (criterio)** a partir del conocimiento de los valores de la **otra variable X (predictora)** con la que sabemos que mantiene un buen nivel de relación.

Estas son las **fórmulas de la regresión lineal**:

$$Y' = a_{yx} + b_{yx} \,.\, X_i$$

en directas

$$Y' = \left(\bar{y} - r_{xy}\frac{s_y}{s_x}\bar{x}\right) + r_{xy}\frac{s_y}{s_x}X_i \quad = \quad r_{xy}\frac{s_y}{s_x}.x + \bar{y}$$

en diferenciales

$(X_i - \bar{x})$

a_{yx} $\quad b_{yx}$

$a_{yx} \neq a_{xy} \rightarrow$ 1º la que se pronostica; 2º la predictora

$Y' \rightarrow$ puntuación directa pronosticada en el criterio

$\bar{y} \rightarrow$ media de las puntuaciones en el criterio

$\bar{x} \rightarrow$ media de las puntuaciones en el test

$r_{xy} \rightarrow$ coeficiente de validez

Sx y Sy $\rightarrow$ desviaciones típicas de las puntuaciones en el test y en el criterio, respectivamente

Xi $\rightarrow$ puntuación directa del sujeto en el test.