COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON
(Guía de clase)

FICHA Nº 20

A. De la Ficha nº13, recordamos las nociones de asociación o correlación y las distintas formas en que una relación puede ser analizada.

✎ Siguiendo la tradición principal de la estadística, aquí hemos definido que existe una relación entre dos variables cuando a través de la aplicación de distintas técnicas rechazamos la hipótesis de que son estadísticamente independientes.

- Una parte de estas técnicas se compone de los denominados coeficientes de asociación o de correlación. La otra parte, incluye diversas pruebas de hipótesis (propias de la inferencia estadística) que resultan relevantes en la medida en que estamos trabajando con muestras en lugar de trabajar directamente con toda la población.

- La hipótesis que es sometida a prueba es la hipótesis nula, que (generalmente) se contrapone a la hipótesis sustantiva.

- La elección de un coeficiente de asociación o correlación dependerá de cuál sea la distribución conjunta esperada entre las dos variables de interés (denominadas por lo general como X y Y).

✎ El análisis de la relación entre dos variables, cuando éstas son intervales o de razón puede hacerse siguiendo alguno o todos de los siguientes cuatro objetivos:

i) La **existencia** (si/no) de una asociación simplemente verificada a través del rechazo de la hipótesis nula que afirma independencia estadística.

ii) La **magnitud** de la asociación Esta noción de magnitud de asociación supone que antes o simultáneamente se ha descartado la hipótesis nula de la

independencia estadística.

iii) El **sentido** de la asociación entre las variables distribuidas conjuntamente. Por ejemplo, afirmando que cuando una se incrementa otra disminuye.

iv) La **forma** de la relación entre las variables examinadas. Si se imagina una representación gráfica, se puede esperar que dos variables puedan tener una relación lineal, curvilineal, exponencial, logística, etc. Este es un punto importante porque algunos coeficientes están diseñados para capturar únicamente relaciones lineales.

✎ En el caso que aquí se tratará, la hipótesis sustantiva del análisis sostiene explícita o implícitamente que existe una **relación lineal** entre las dos variables de interés. En tal caso, se podrá aplicar el coeficiente de correlación de Pearson, o coeficiente de correlación producto-momento.

- El principio de isomorfía destacado desde el comienzo se visualizará muy claramente aquí: sólo si la estructura de la hipótesis sustantiva supone una relación lineal, tendrá sentido utilizar el coeficiente de Pearson.

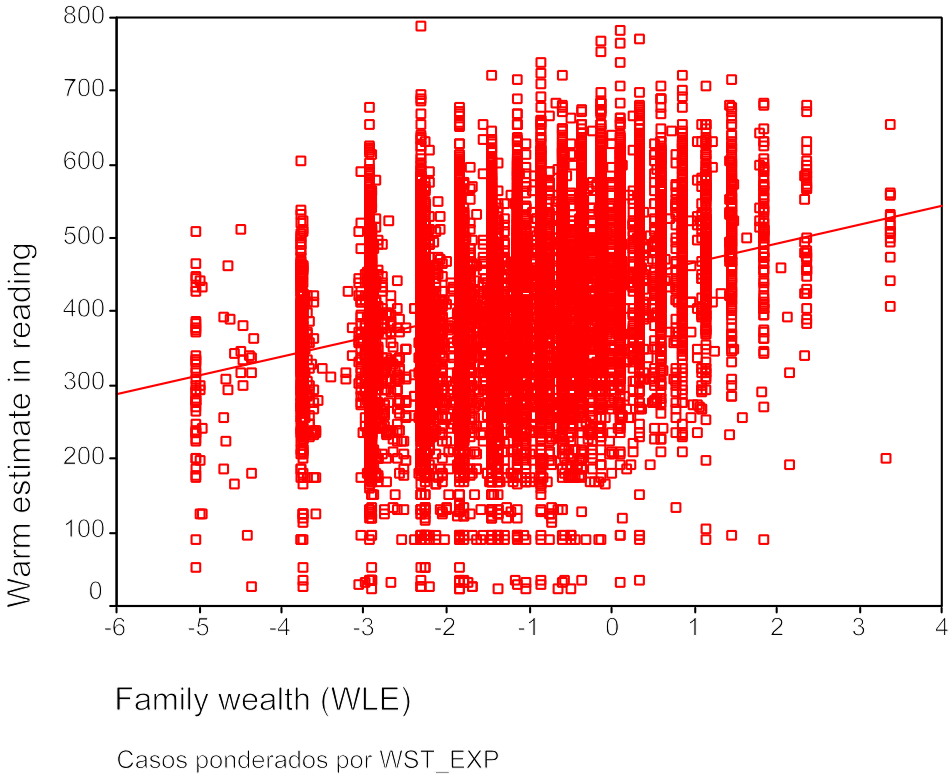
- Si la relación hipóteticamente se supone no lineal, **no deberá utilizarse este coeficiente para contrastar la hipótesis.**

- Si el coeficiente de Pearson calculado para la distribución conjunta informa que no existe relación, **deberá tenerse muy presente de que la conclusión es que No hay relación lineal.**

- Es conveniente en estos casos, comenzar el análisis graficando la relación entre X y Y mediante un diagrama de dispersión ().

COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON
(Guía de clase)

B. Siguiendo la hipótesis de que un mayor capital económico permite a los hogares diversificar y enriquecer las alternativas, los estímulos y los recursos educativos de sus hijos adolescentes y que estos a su vez, elevan el nivel de dominio de la comprensión lectora, se ha trazado el siguiente gráfico para los cinco países de América Latina (Argentina, Brasil, Chile, México y Perú) participantes del ciclo 2000 del Programme for International Student Assessment (PISA).



- Se puede apreciar que empíricamente parecería existir evidencia para sostener la validez de la hipótesis enunciada más arriba. A mayores valores

---

del índice de posesiones y bienestar ("wealth") de la familia (variable tipificada), mayores niveles de desempeño en lectura.

- En consecuencia, parece razonable dar el siguiente paso y construir una medida que permita informar de la magnitud de la relación lineal entre las variables.

C. El coeficiente de Pearson es un estadístico muy potente:

✎ Tiene la propiedad de controlar a la vez tres aspectos importantes:

- Controla el tamaño de la muestra con que se está trabajando: lo cual implica que la fuerza de la relación no varía según el N bajo análisis.

- Controla la métrica de las variables analizadas: lo cual implica que si por ejemplo, en lugar de expresar el indicador socioeconómico de bienestar en puntajes Z lo expresara en dólares, el valor del coeficiente no se alterará.

- Encierra el rango en que varía el coeficiente: lo cual implica que implica que una relación lineal perfecta y positiva entregará un coeficiente de +1 y una relación lineal perfecta pero inversa entregará un coeficiente de - 1.

✎ Ha sido desarrollado sobre una base lógico-estadístico desde la cual es posible permite entender la fórmula general. En este apartado seguimos lo planteado por Cortés (2000: 169-175).

- Para analizar la relación entre dos variables, un primer paso intuitivo podría consistir en un estadístico que fuera el resultado de sumar el producto de los valores en cada unidad de las variables de interés X,Y. Esta suma tendría N términos:

[1]

$$E = \sum (X_i Y_i)$$

- Si las dos variables tuvieran una relación positiva entre ellas, el incremento de X conllevaría el incremento en los valores de Y. En consecuencia, el producto sería positivo. Si por el contrario, la relación fuera lineal pero inversa, el producto sería negativo y la suma daría un valor negativo indicando esta situación.

- Frente a este primer paso, surgiría inmediatamente la crítica: es evidente que con sólo agregar un número adicional de casos, la fuerza de la relación se incrementaría. Para resolver esta sensibilidad del estadístico E al tamaño de la muestra N, se propone un segundo paso consistente en la normalización:

[2]
$$E = \frac{\sum (X_i Y_i)}{N}$$

- Sin embargo, el estadístico aún es susceptible de otra crítica: si se modifica la métrica de las variables de interés, por ejemplo pasando de puntajes Z a dólares, por este mero cambio el valor del estadístico se incrementaría.

- Se propone por lo tanto, controlar la métrica de las variables de interés, expresando los valores de X y de Y en términos de desviaciones respecto a cada una de las respectivas medias aritméticas.

[3]
$$Cov_{(X,Y)} = \frac{\sum [(X_i - \overline{X})(Y_i - \overline{Y})]}{N}$$

- El estadístico que se presenta en la ecuación [3] se conoce con el nombre de **covarianza** (Cov).

- Sin embargo, el coeficiente sigue presentando el problema de que su valor no queda encerrado en un rango fijo y predeterminado, idealmente entre -1 y +1. En consecuencia de propone una última transformación:

$$r_{(X,Y)} = \frac{\sum [(X_i - \overline{X})(Y_i - \overline{Y})] / N}{\sqrt{[(X_i - \overline{X})^2 / N][(Y_i - \overline{Y})^2 / N]}}$$

Simplificando:

$$r_{(X,Y)} = \frac{\sum [(X_i - \overline{X})(Y_i - \overline{Y})]}{\sqrt{[(X_i - \overline{X})^2][(Y_i - \overline{Y})^2]}}$$

O:

$$r_{(X,Y)} = \frac{Cov_{X,Y}}{S_X * S_Y}$$

[4]

✎ El siguiente cuadro muestra los distintos pasos del cálculo hechos, con los pasos intermedios y los valores de las sucesivas expresiones colocados entre paréntesis.

- Se puede apreciar que para la distribución de valores aquí incluida, el coeficiente de correlación alcanza el valor de 0.78, lo cual puede considerarse entre fuerte y muy fuerte.

- Se han introducido en el mismo cuadro, otras tres columnas donde se hace el ejercicio de multiplicar los valores del índice de bienestar por 10 para mostrar que los estadísticos propuestos en los pasos [1] a [3] se alteran pero que el coeficiente de correlación de Pearson mantiene su mismo valor.

COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON (Guía de clase)

Cuadro 20.1
Ejemplificación del desarrollo del coeficiente de correlación

	Nivel de bienestar Económico (media = 0)	Puntaje en Lengua (media = 500)	X*Y [1]	$(X_i - \bar{x})$	$(Y_i - Y)$	$(X_i - \bar{x})^* (Y_i - Y)$ [3]	Coeficiente de correlación [4]	Bienestar * 10	[1] X*Y	[3]	[4]
1	-1.60	341.72	-546.75	-1.60	-158.28	253.25		-16.00	-5467.52	2532.48	
2	-0.13	438.95	-57.06	-0.13	-61.05	7.94		-1.30	-570.64	79.37	
3	0.54	553.12	298.68	0.54	53.12	28.68		5.40	2986.85	286.85	
4	-0.13	425.54	-55.32	-0.13	-74.46	9.68		-1.30	-553.20	96.80	
5	0.34	508.92	173.03	0.34	8.92	3.03		3.40	1730.33	30.33	
6	-0.36	415.83	-149.70	-0.36	-84.17	30.30		-3.60	-1496.99	303.01	
7	2.31	592.56	1368.81	2.31	92.56	213.81		23.10	13688.14	2138.14	
8	-0.60	389.47	-233.68	-0.60	-110.53	66.32		-6.00	-2336.82	663.18	
9	1.14	546.62	623.15	1.14	46.62	53.15		11.40	6231.47	531.47	
10	1.62	557.25	902.75	1.62	57.25	92.75		16.20	9027.45	927.45	
$\sum$	3.13	4769.98	2323.91			758.91		31.30		7589.07	
$\bar{x}$	0.31	477.00	232.39			75.89		3.13		758.91	
S^2	1.31	7250.46						130.79			
S	1.14	85.15						11.44			
r							0.78				0.78

COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON
(Guía de clase)

D. Una vez clarificado el proceso de desarrollo del coeficiente, conviene reseñar algunas de sus propiedades, para lo cual seguimos a Gujarati (2004:82). Algunas ideas han sido planteadas ya pero conviene reafirmarlas una vez más:

- El coeficiente puede tener signo positivo o negativo, dependiendo del signo del término en el numerador de [4], es decir del producto de las desviaciones de X e Y respecto a sus respectivas medias.

- Su rango queda limitado entre -1 y + 1, es decir:

$$-1 \leq r \leq +1$$

- Es un coeficiente simétrico por formulación:

$$r_{(X,Y)} = r_{(Y,X)}$$

- Es independiente del origen o de la métrica en la que estén expresadas las variables.

- Si X e Y son estadísticamente independientes, el coeficiente de correlación entre ellos será igual a cero. **Pero**, un coeficiente igual a cero **no necesariamente está indicando que exista independencia entre las dos variables.**

- Lo anterior se debe a que es un estadístico de asociación **lineal**; su uso en relaciones no lineales no tiene sentido.

Si:

$Y = X^2$

Entonces:

$r_{(X,Y)} = 0$
