22

Análisis de Correlación

En este capítulo se continúa con el estudio de estadísticas en dos dimensiones. Cuando se trata de medir el *grado de asociación entre dos magnitudes* biológicas cualquiera se usa el Análisis de Correlación, mientras que cuando se trata de la relación funcional entre ambas se usa el Análisis de Regresión, como se vio en el capítulo anterior. Aquí, se comienza comparando los conceptos de regresión y correlación para ver cuando se aplica uno u otro modelo. Luego se explica la fórmula del producto-momento de Pearson y sus relaciones matemáticas básicas. Los cálculos básicos se ejemplifican para muestras pequeñas y medianas. Se plantean los tests de significación en correlación más comunes. Y por último se muestran algunas aplicaciones básicas de la correlación.

22.1 Introducción

No siempre resulta sencillo saber cual de los dos métodos se debe emplear en el estudio de un problema dado. Ha habido confusión en la literatura y entre los investigadores, al respecto. Por eso, se insistirá una vez más en tratar de distinguir claramente entre ambos casos. Es muy frecuente encontrar casos de correlación tratados como una regresión y viceversa. Hay varias razones para ello:

- Las relaciones matemáticas entre ambos modelos son muy estrechas, se puede pasar de uno a otro con mucha facilidad en los cálculos y eso siempre ha sido una tentación muy fuerte. Básicamente, el cuadrado del coeficiente de correlación es el cociente entre la variación explicadas y la total, que se calculan exactamente igual en ambos modelos:

$r = \frac{(\sum xy)}{\sqrt{\sum x^2 \sum y^2}}$ es el *coeficiente de correlación* (fórmula del producto-momento de K.R. Pearson)

$b = \frac{\sum xy}{\sum x^2}$ es el coeficiente de regresión lineal

Y la relación entre ambos es: $r^2 = b \frac{\sum xy}{\sum y^2}$

Esta última relación no tiene ningún significado conceptual, simplemente se trata de una analogía en los cálculos. Desdichadamente, muchos la usan como si fuera lo mismo y mezclan la regresión con la correlación, lo cual no es correcto.

- En los textos antiguos nunca se hizo una distinción lo suficientemente clara entre los dos conceptos. Aún hoy, esto no está totalmente superado, sobretodo por las reimpresiones actuales de los textos clásicos escritos en las primeras décadas del siglo pasado. Hay autores que usan ambos términos como si fuesen sinónimos, lo que aumenta más la confusión.

- Los métodos empíricos desarrollados para ingeniería y otras disciplinas, se usan para simplificar las relaciones reales a términos prácticos. Por ejemplo, las curvas de rendimiento de motores eléctricos, de máquinas térmicas, etc. En esos casos, no interesa la relación entre ambas magnitudes sino obtener una gráfica empírica para el uso diario.

- A veces, aunque el método escogido es el correcto, los datos disponibles no permiten aplicarlo.

Se puede revisar este tema desde el punto de vista del investigador. Tomando por caso un experimento donde se desea establecer el contenido de colesterol en la sangre humana como función del peso corporal. Para independizarse del factor edad y sexo se escoge en forma aleatoria un grupo de personas del mismo sexo y edad, midiendo en cada una ambos valores. Con el grupo de valores así obtenido se puede calcular la regresión entre ambas magnitudes. Sin embargo, el investigador no tiene bajo su control a ninguna de las variables; por lo tanto, las condiciones básicas de un Modelo I de regresión no se cumplen porque ambas variables han sido medidas con error. Casos como este abundan en la bibliografía a pesar de no ser legítimos. Se podría pensar que se trata de un Modelo II de regresión, pero salvo casos muy especiales como el de Berkson, esto tampoco es correcto. En un Modelo II la relación entre ambas variables puede escribirse con la ecuación:

$$\mu y = \alpha + \beta \ \mu x$$

El verdadero valor de Y es igual al valor poblacional de α, más valor poblacional del coeficiente de regresión multiplicado por el verdadero valor de X. Pero como ambas variable se miden con error, será $Y = \mu y + \varepsilon y$ (donde εy es el término de error expresado con una distribución normal) y $X = \mu x + \varepsilon x$ (donde εx es el término de error de la variable X).
Luego, la covarianza de ambas sería:

$$\sum XY = \sum (\mu x + \varepsilon x)(\mu y + \varepsilon y) = \sum [(\mu x \, \mu y) + (\mu y \, \varepsilon x) + (\mu x \, \varepsilon y) + (\varepsilon x \, \varepsilon y)] \qquad \mu x$$

Donde se espera que, salvo el primer término, todos los demás se anulen pues las partes debidas al error son independientes entre sí y además de ambos valores esperados. Sin embargo, las desviaciones $Di = Yi - Y^*i$ no son ahora independientes de Y^*i, por lo que se invalidan los tests de significación convencionales. El caso especial de Berkson, es cuando las variables independientes se miden con error, aunque son "controladas" por el investigador que posee información acerca de las respectivas varianzas y sabe que los valores X y εx no están correlacionados.

En conclusión, lo correcto en el caso del experimento de nivel de colesterol y el peso corporal no es hacer una regresión, sino un análisis de correlación entre ambas variables, para el cual esos datos son convenientes, si lo que se busca es alguna ecuación que describa la dependencia de Y (colesterol) sobre el peso (X). El caso contrario es cuando se quiere hallar un coeficiente de correlación, cuando los datos se han calculado apropiadamente como de regresión. Por ejemplo: medir los latidos del corazón en función de la temperatura ambiente. En este caso, el investigador puede decir que los valores de temperatura fueron elegidos al azar. Pero en el fondo se sabe que tales valores son "controlables" por el investigador mediante un adecuado sistema de frío. Se somete al sujeto a una cierta temperatura (elegida al azar o no) regulada por un termostato y se miden los latidos. O sea, se puede calcular el coeficiente de correlación con estos datos, pero solo sería un valor numérico en vez de un valor paramétrico de la correlación. A pesar, de que se puede relacionar el cuadrado de este coeficiente con el cociente entre la variación explicada y la total, no es de ninguna manera una indicación de correlación paramétrica. Se pueden resumir estos conceptos en un cuadro como el siguiente:

Cuadro 22.1: Regresión versus Correlación.

El investigador quiere:	Y aleatoria, X fija	X e Y aleatorias
Establecer y estimar la *dependencia* de una variable sobre la otra.	Modelo I de Regresión.	Modelo II de Regresión. (caso especial de Berkson)
Establecer y estimar el grado de *asociación* entre ambas variables	No es significativa.	Análisis de Correlación. (test de significación válido solo si X e Y son variables normales bidimensionales)

22.2 Fórmula del producto-momento

La forma más general de esta fórmula para obtener el *coeficiente de correlación* **r** es:

$$r = \frac{Sxy}{DSx\ DSy}$$

Dónde:

Sxy : es la covarianza entre ambas variables.
DSx : es el desvío estándar de una variable cualquiera X.
DSy : es el desvío estándar de otra variable cualquiera Y.

$$Sxy = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{(n-1)} = \frac{\sum x\ y}{(n-1)}$$

$$DSx = \sqrt{\frac{\sum \left(X - \overline{X}\right)^2}{n-1}} = \sqrt{\frac{\sum x^2}{n-1}} \qquad \text{Análogamente, } DSy = \sqrt{\frac{\sum y^2}{n-1}}$$

Reemplazando y simplificando la fórmula del producto momento queda:

$$r = \frac{\sum x\,y}{\sqrt{\sum x^2 \sum y^2}} \quad : \textit{coeficiente de correlación}$$

Se puede reagrupar esta ecuación de la manera siguiente:

$$r^2 = \frac{\left(\sum xy\right)^2}{\sum x^2 \sum y^2} = \frac{\left(\sum xy\right)^2}{\sum x^2} \cdot \frac{1}{\sum y^2} = \sum (x^*)^2 \cdot \frac{1}{\sum y^2} = \frac{\sum (X^* - \overline{X})^2}{\sum (Y - \overline{Y})^2}$$

$$r^2 = \frac{\text{Variación Explicada en X}}{\text{Variación Total en Y}}$$ (ver punto 21.3 del capítulo anterior)

En una correlación, el cuadrado del coeficiente de correlación es el cociente entre la Variación Explicada de una de las variables y la Variación Total de la otra. Este cociente, se denomina *coeficiente de determinación* y puede variar entre cero y uno. En un caso extremo, cuando la covarianza entre ambas variables no existe, valdrá cero y se piensa que no hay ningún grado de correlación entre ambas. En el otro extremo, cuando el grado de asociación entre ambas variables es perfecto valdrá uno. Por su parte, el coeficiente de correlación variará entre –1 y +1.

22.3 Cálculo del coeficiente de correlación

La manera más sencilla es resolviendo el siguiente ejemplo. En una investigación se eligieron al azar nueve individuos de aproximadamente 30 años, de una misma ciudad, considerados sanos. A cada uno de ellos se le midió el peso y el nivel de colesterol en sangre. Los resultados se muestran a continuación. Hallar el coeficiente de correlación entre ambas variables.

	Colesterol	Peso	x =	y =	x^2	y^2	x.y
	X	**Y**	**X-214**	**Y-73**			
	210	70,2	-4	-2,8	16	7,84	11,2
	122	62,4	-92	-10,6	8464	112,36	975,2
	309	95,4	95	22,4	9025	501,76	2128
	198	68,9	-16	-4,1	256	16,81	65,6
	260	75,2	46	2,2	2116	4,84	101,2
	230	76	16	3	256	9	48
	175	64,5	-39	-8,5	1521	72,25	331,5
	198	64,2	-16	-8,8	256	77,44	140,8
	224	80,2	10	7,2	100	51,84	72
Total	**1926**	**657**	**0**	**0**	**22010**	**854,14**	**3873,5**
Media	**214**	**73**					

r^2 = 0,7981013

Paso 1) Se vuelcan los datos a una tabla como la anterior y se calculan los totales y promedios respectivos.

Paso 2) Se calculan las diferencias de cada dato respecto a su media y se vuelcan los resultados en las columnas tercera y cuarta.

Paso 3) Los valores obtenidos en el paso anterior se elevan al cuadrado y se colocan en las columnas quinta y sexta. Y en la última columna se coloca el producto de los valores de tercer y cuarta columna.

Paso 4) Los totales de las tres últimas columnas se usan para el cálculo del coeficiente:

$$\sum x^2 = 22010 \quad ; \quad \sum y^2 = 854{,}14 \quad ; \quad \sum x\,y = 3873{,}5$$

Luego $r = \dfrac{3873{,}5}{\sqrt{(22010)(854{,}14)}} = 0{,}8933$

22.4 Tests de significación en correlación

Hay varias maneras de efectuar los tests de significación en correlación, de acuerdo a la hipótesis nula que se plantee. Básicamente se agrupan en cuatro casos:

. H_0 : $\rho = 0$ No hay correlación. El verdadero valor de r es nulo.
. H_0 : $\rho = r$ El coeficiente de correlación vale r.
. H_0 : $\rho_1 = \rho_2$ No hay diferencias entre dos coeficientes de correlación.
. H_0 : $\rho_1 = \rho_2 = \rho_3 = ... = \rho_k$ Todas las muestras provienen de la misma población.

Caso 1) Cuando se postule $H_0 : \rho = 0$, se tiene el caso más sencillo en la correlación de dos variables. El planteo supone que no existe la correlación entre ambas variables y hay tres maneras de proceder. La primera es usando la Tabla 20 del Anexo donde en la primer columna, para un número k = 1, y los grados de libertad obtenidos con $\nu = n - 2$, se obtienen los valores críticos para 95% y 99% de confianza. La segunda forma es usando el test clásico con el modelo Student. Y la tercera, cuando las muestras son grandes (n > 50), con la transformación Z de Fisher, estadígrafo con una distribución aproximación normal.

Ejemplo 1) Usando los datos del problema visto en el punto anterior determinar si los resultados permiten validar que hay correlación.
Para esto, se plantea una hipótesis a ser rechazada para tener la validación buscada. Esto es:

H_0 : $\rho = 0$ No hay correlación. El verdadero valor de r es nulo.
H_0 : $\rho \neq 0$ Hay correlación. El verdadero valor de r es diferente de cero.

Método 1) Se busca en la primer columna de la Tabla 20 del Anexo, con $\nu = n - 2 = 7$ grados de libertad y resulta: $R_{0.95;\,7} = 0{,}666$ y $R_{0.99;\,7} = 0{,}798$

Entonces $r = 0{,}8933^{**}$ y se rechaza la hipótesis nula. Por lo tanto, se tiene evidencia muy significativa de correlación.

Método 2) Se usa el modelo Student de dos colas, comparando con tablas al estadígrafo $t_{\alpha;7=n-2}$:

$$t = \frac{(r - 0)}{\sqrt{\frac{(1 - r^2)}{(n - 2)}}} = r \,.\, \sqrt{(n - 2)/(1 - r^2)} = 0{,}8933 \,.\, \sqrt{(9 - 2)/(1 - 0{,}8933^2)} = 5{,}26^{**} > t_{0{,}99;7}$$

La conclusión es similar a la vista con el método anterior.

Ejemplo 2) Phillips en 1929 midió las longitudes de la vena transversal en las alas derecha e izquierda de 500 abejas obreras y obtuvo un coeficiente de correlación r = 0,837. Determinar si el coeficiente de correlación es significativo.

$H_0 : \rho = 0$ No hay correlación. El verdadero valor de r es nulo.
$H_0 : \rho \neq 0$ Hay correlación. El verdadero valor de r es diferente de cero.

Se emplea el estadígrafo t* y se lo compara con la t de Student para infinitos grados de libertad. O sea, se trata de una distribución gaussiana y por ese motivo se prefiere llamarlo z = t*. Donde Z se calcula con la transformación de Fisher y el desvío estándar poblacional es la inversa de la raíz cuadrada del número de mediciones, menos 3. Esto es:

$$t^* = z = \frac{(Z - 0)}{\frac{1}{\sqrt{n - 3}}} = Z\,\sqrt{(n - 3)}$$

donde $Z = 0{,}5 \ln [\frac{1 + r}{1 - r}] = 0{,}5 \ln [11{,}27] = 1{,}2111$

$z = 1{,}2111\ (\ 22{,}29\ ) = 27^{***} >>> z_{0{,}001} = z_{,\alpha}$ de la función de Gauss

Aplicando la fórmula gaussiana al valor z = 27 se obtiene una probabilidad muy pequeña (del orden de 10^{-6}) lo que permite rechazar la hipótesis nula con mucha evidencia significativa.

Caso 2) El otro caso es $H_0 : \rho = \mathbf{a}$ la hipótesis supone un valor cualquiera **a** para el coeficiente de correlación poblacional. Aquí se emplea otra vez el modelo de Fisher, para obtener un estadígrafo z con una distribución gaussiana, para muestras grandes (n>50) con:

$$z = \frac{Z - \mu_z}{\sigma_z} \quad \text{Donde} \quad \mu_z = 0{,}5 \ln [\frac{1 + a}{1 - a}] \quad \text{y} \quad SE(z) = \sigma_z = \frac{1}{\sqrt{n - 3}}$$

Si la muestra es más pequeña $10 < n < 50$ se debe efectuar una corrección.

Ejemplo 3) Con los datos del problema anterior se desea comprobar sí:

H_0 : $\rho = 0,5$ En este caso se emplea la transformación Z de Fisher con:
H_1 : $\rho \neq 0,5$

$$z = \frac{Z - \mu_z}{\sigma_z}$$

Donde:

$$\mu_z = 0,5 \ln \frac{(1+\rho)}{(1-\rho)} = 0,5 \ln (1,5/0,5) = 0,5493$$

$$\sigma_z = 1 / \sqrt{(n-3)} = SE(z) = 0,045$$

$$Z = 0,5 \ln \frac{(1+r)}{(1-r)} = 0,5 \ln (11,27) = 1,2111$$

Por lo tanto:

$z = (1,2111 - 0,5493) / 0,045 = 17,71^{***}$ (versus el valor gaussiano $z_{,\alpha}$)

Se rechaza la hipótesis nula con valores altamente significativos. La probabilidad de que $\rho = 0,5$ es muy pequeña.

Ejemplo 4) Se ha tomado una muestra de n = 12 pares de datos de una población y se obtuvo un coeficiente de correlación r = 0,8652. Se desea saber si el coeficiente poblacional es igual a 0,8.

H_0 : $\rho = 0,8$ En este caso se emplea la transformación Z de Fisher con la corrección
H_1 : $\rho \neq 0,8$ de Hotelling:

$$z = \frac{Z^* - \mu^*_z}{\sigma^*_z}$$

Donde:

$$Z = 0,5 \ln \frac{(1+r)}{(1-r)} = 0,5 \ln (1,8652/0,1348) = 1,3137$$

$$\mu_z = 0,5 \ln \frac{(1+\rho)}{(1-\rho)} = 0,5 \ln (1,8/0,2) = 1,0986$$

Pero ahora se debe corregir con:

$$Z^* = Z - [(3Z + r) / (4n - 4)] = 1,3137 - [(4,8062) / 44] = 1,2045$$

$$\mu_{*z} = \mu_z - [(3\mu_z + \rho) / (4n)] = 1,0986 - [(4,0958)/48] = 1,0133$$

$$\sigma_{*z} = SE(z) = 1/\sqrt{n-1} = 1/\sqrt{11} = 0,3015$$

Luego es: $z = \dfrac{Z^* - \mu^*_z}{\sigma^*_z} = (1{,}2045 - 1{,}0133) / (0{,}3015) = 0{,}634$ (no significativo)

Con este resultado no se puede rechazar la hipótesis nula.

Caso 3) Ahora es $H_0 : \rho_1 = \rho_2$ Se supone que hay dos muestras tomadas de la misma población y tienen el mismo coeficiente de correlación poblacional. Se usa la transformación de Fisher para la comparación de ambos coeficientes de correlación, en forma similar a la tratada en el modelo gaussiano para diferencia de medias muestrales independientes. Esto es el estadígrafo:

$z = \dfrac{Z_1 - Z_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}}$ se distribuye normalmente y se puede comparar con el de tablas.

Ejemplo 5) Se desea determinar si hay diferencias significativas entre dos coeficientes de correlación $r_1 = 0{,}8$ y $r_2 = 0{,}5$, calculados con muestras de tamaños 50 y 60 respectivamente.

$H_0 : \rho_1 = \rho_2$
$H_1 : \rho_1 \neq \rho_2$

$Z_1 = 0{,}5 \ln \dfrac{(1 + r_1)}{(1 - r_1)} = 0{,}5 \ln (1{,}8/0{,}2) = 1{,}0986$

$Z_2 = 0{,}5 \ln \dfrac{(1 + r_2)}{(1 - r_2)} = 0{,}5 \ln (1{,}5/0{,}5) = 0{,}5493$

Luego es:

$$z = (1{,}0986 - 0{,}5493) / \sqrt{(1/47) + (1/57)} = 0{,}5493 / 0{,}197 = 2{,}79^{**}$$

Se rechaza la hipótesis nula con resultados muy significativos.

22.5 Comparación entre dos o más coeficientes

Hay oportunidades en que se miden las dos magnitudes clínicas en varias muestras. Se obtienen de esta forma varios coeficientes de correlación. Esto es, como si se tuviesen varias mediciones del mismo índice. Lo que se desea averiguar es si todos ellos provienen de la misma población cuyo verdadero coeficiente de correlación es $\Xi (r) = \rho$. Puede ser aplicado en casos donde se toma en cuenta la raza, la zona geográfica donde habitan, la edad, el sexo, etc. Una manera de decidir si los coeficientes de correlación hallados son homogéneos entre sí, es encontrar el valor poblacional con una estimación a través de todas las muestras y realizar una validación a través del Modelo de Fisher para correlación. Se puede hacer tomando una suma de cuadrados ponderada del los valores de los coeficientes r, transformados con Z, la cual se distribuye con una distribución Chi cuadrado con N -1 grados de libertad.

Para ilustrar el procedimiento se ha elegido el siguiente caso:

Ejemplo) En un estudio del Ministerio de Salud Pública, en diez localidades de la provincia de Misiones, se tomaron muestras al azar de diferentes tamaños de su población y a cada individuo seleccionado se le midió peso y talla. El sexo se repartió mitad y mitad en cada grupo. Las edades corresponden a individuos que se hallan cursando la escuela primaria. En la Tabla 24.1 siguiente:

Tabla 24.1: Correlación entre talla y peso en 10 localidades misioneras.

	n	n-3	r	Z	Z^2	(n-3)Z	$(n-3)Z^2$
	200	197	0,42	0,4477	0,20044	88,1969	39,4858
	150	147	0,45	0,4847	0,23493	71,2509	34,5353
	140	137	0,49	0,5361	0,2874	73,4457	39,3742
	130	127	0,48	0,523	0,27353	66,421	34,7382
	90	87	0,57	0,6475	0,41926	56,3325	36,4753
	120	117	0,51	0,5627	0,31663	65,8359	37,0459
	160	157	0,47	0,5101	0,2602	80,0857	40,8517
	50	47	0,55	0,6184	0,38242	29,0648	17,9737
	160	157	0,49	0,5361	0,2874	84,1677	45,1223
	170	167	0,61	0,7089	0,50254	118,3863	83,924
Total:	**1370**	**1340**				**733,1874**	**409,5264**

En la primer columna se colocan los respectivos tamaños muestrales de cada localidad. En la tercer columna se vuelcan los respectivos coeficientes de correlación encontrados. Los pasos a seguir son:

Paso 1) Se calculan los tamaños muestrales menos tres y se colocan en la segunda columna.

Paso 2) Se obtienen los valores de r transformados con la relación Z:
Por ejemplo, para la primer localidad será:

$$Z_1 = 0{,}5 \ \ln \frac{(1+r_1)}{(1-r_1)} = 0{,}5 \ln (1{,}42 / 0{,}58) = 0{,}4477$$

Paso 3) Una vez calculados todos los valores de Z se vuelcan en la cuarta columna, y sus cuadrados se colocan en la quinta columna. O sea, en la quinta columna se colocan los resultados de la cuarta multiplicados entre sí.

Paso 4) Luego a cada valor de estas columnas se lo multiplica por su tamaño muestral respectivo y los resultados se colocan en las dos últimas. O sea, (n-3) se multiplica por la cuarta columna y el resultado se coloca en la sexta. Lo mismo con (n-3) por la quinta y se coloca en la séptima.

Paso 5) Se obtiene el tamaño muestral corregido T con:

$$N = \sum Ni = 200 + 150 + ... + 170 = 1.370$$

$$T = \sum (Ni - 3) = (200-3) + (150 - 3) + ... + (170 - 3) = 1.340$$

O bien, T = N – 3 (10) = 1.370 – 30 = 1.340

Paso 6) Se calcula el promedio de Z con:

$$\Xi (Z) = \frac{\sum (Ni - 3)\, Zi}{\sum (Ni - 3)} = \frac{\text{Total de columna sexta}}{\text{Total de columna segunda}} = (733,1874) / 1340 = 0,54715$$

El valor promedio de Z ponderado es: $\Xi (Z) = 0,547155$

Paso 7) Se calcula la suma de cuadrados de Z ponderada con:

$SSz = \sum (Ni - 3)\, Z^2_i$ = Total de la columna séptima = 409,5264

Paso 8) Se calcula el término de corrección:

$TC = \Xi (Z) . \sum (Ni - 3)\, Z_i$ = *Paso 6* . Total de la columna sexta = (0,547155) . (733,1874)

TC = 401,167

Paso 9) Se efectúa el test de homogeneidad:

Ho : $r_i = r_2 = ... = r_k$ Todos los coeficientes provienen de la misma población.
H1 : Los r_i son diferentes y no provienen de la misma población.

El estadígrafo de comparación es:

$$X^2 = SSz - TC = \sum (Ni - 3)\, Z^2_i - \Xi (Z) . \sum (Ni - 3)\, Z_i = 409,5264 - 401,1677 = 8,359$$

La comparación es:

$$X^2 = 8,3587 < \chi^2_{0,95\,;\,9} = 16,919$$

No hay suficiente evidencia como para rechazar la hipótesis nula.

Cálculo del coeficiente de correlación poblacional.

Paso 10) Con el valor promedio de Z, se calcula la mejor estimación de ρ:

$$\Xi (r) = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1} = \frac{e^{2(0,547155)} - 1}{e^{2(0,547155)} + 1} = 0,4984 \approx \rho$$ (Notar que no es el promedio ponderado de los r).

22.6 Modelo no paramétrico de Kendall

Hay ocasiones en que se sabe que los datos no se distribuyen mediante una distribución normal bi-variante. Otras veces las magnitudes medidas no son de tipo continuo sino cualitativas de tipo ordinal, que permitan comparar a las dos muestras, tales como puntajes, etc. En todos estos casos no se pueden emplear los modelos vistos hasta ahora y se requiere de un modelo no paramétrico equivalente como el de Kendall, para una correlación ordenada. Hay otros modelos para distintas variantes; el lector interesado los puede encontrar en la obra de Siegel indicada en la bibliografía. La condición es que los datos se puedan medir en por lo menos una escala ordinal

El ejemplo típico es como se ordenan a los alumnos de un curso de Bioestadística por sus calificaciones de parciales y de concepto. Por ejemplo, se puede afirmar que A es el mejor alumno, B es el segundo mejor estudiante, que C y D son iguales entre sí pero no son tan buenos como B y así sucesivamente. En otra materia con los mismos alumnos, otro profesor puede ordenarlos según su criterio, respecto a como se comportan en sus clases. Se supone que los dos grupos de valores este correlacionados entre sí porque los alumnos son los mismos, aunque sometidos a diferentes exigencias. Sin embargo, para probar que esto ocurre se necesita un test de validación de correlación ordenada como el de Kendall. En una Farmacia se pueden ordenar los diferentes artículos de acuerdo a la cantidad vendida de cada uno y construir una tabla ordinal con los mejores del ranking. Lo mismo se puede repetir en una sucursal de la misma farmacia. Se espera que haya una correlación entre ambos listados con la ubicación ordenada de los artículos, y otra vez el modelo de Kendall puede aplicarse para verificar si esto es cierto. En un laboratorio de Análisis Clínicos se puede desarrollar la muestra extraída a un paciente en un caldo de cultivo apropiado. Luego de hacer el antibiograma, se pueden ordenar los resultados de acuerdo al diámetro de la aureola medido en la caja de Petri. Con la muestra de otro paciente que tenga la misma enfermedad se procede en forma similar, y se espera que los dos grupos de datos ordenados guarden alguna correlación. Para ilustrar el procedimiento se simulan datos para un problema como el de los antibiogramas, para 15 pacientes, midiendo el diámetro de la halo de inhibición.

N	Y1	R1	Y2	R2
1	8,7	8	5,95	9
2	8,5	6	5,65	4
3	9,4	9	6	10
4	10,0	10	5,7	6,5
5	6,3	1	4,7	2
6	7,8	5	5,5	3
7	11,9	15	6,4	15
8	6,5	2	4,18	1
9	6,6	3	6,15	13
10	10,6	12	5,93	8
11	10,2	11	5,7	6,5
12	7,2	4	5,68	5
13	8,6	7	6,13	12
14	11,1	13	6,3	14
15	11,6	14	6,03	11

A la muestra de cada paciente se la siembra en un caldo de cultivo apropiado. Se prueban dos antibióticos 1 y 2. Los diámetros de las aureolas resultantes se colocan en la segunda y cuarta columna, bajo Y1 e Y2 respectivamente. Luego se calculan los rangos que corresponden a cada muestra y se colocan en las columnas R1 y R2.

Así, el dato más pequeño de la muestra 1 es 6,3 y le corresponde el rango 1, el segundo con rango 2 es 6,5 y así sucesivamente hasta el más grande 11,9 con rango 15.

En la otra serie de datos se procede igual. Como hay un empate entre el rango 6 y 7, se coloca el rango promedio a ambos (Y = 5,7).

El coeficiente de correlación ordenada de Kendall es un estadígrafo y no un parámetro, pero generalmente se lo simboliza con la letra griega tau (τ). La fórmula para este estadígrafo es:

Si no hay empates $\tau = O / N(N-1)$

Si hay empates $\tau = \dfrac{O}{\sqrt{\left[N(N-1) - \sum E_2\right]\left[N(N-1) - \sum E_1\right]}}$

N es el tamaño muestral, en este caso es N = 15

O es una cantidad calculada en función del orden, que se puede determinar de varias formas.

$\sum E_1$: Es la suma de las cantidades empatadas en Y1.

$\sum E2$: Es la suma de las cantidades empatadas en Y2.

La idea es que si la variable Y2 está perfectamente correlacionada con Y1, entonces deberían tener los mismos rangos, es decir, ordenadas de igual manera. Sin embargo, si la correlación es menor el orden de los Y2 no se corresponderá con los de Y1. La cantidad O mide el grado en que la *segunda* variable corresponde al orden de la primera. Su valor máximo será: N (N-1).

A continuación se explican los pasos a seguir para resolver el problema:

Paso 1) Si una de las variables, como en este caso, tiene empates, se ordenan los pares según la variable que no los tenga. Si ambas tienen empates, entonces puede ser cualquiera. A continuación del ordenamiento según la primera variable, se colocan los correspondientes rangos de la otra. Para cada caso hay que calcular el número de rangos de la segunda más grande. Comenzando con R1 = 1, le corresponde R2 = 2. Salvo el R2 =1, todos los demás son mayores, entonces habrá 13 rangos R2 mayores que R2 =2 y así sucesivamente. Como se muestra a continuación:

R1	R2
1	2
2	1
3	13
4	5
5	3
6	4
7	12
8	9
9	10
10	6,5
11	6,5
12	8
13	14
14	11
15	15

Rangos que son superiores al R2 en cada caso:
13; 5; 3; 4; 12; 9; 10; 6,5; 6,5; 8; 14; 11; 15. O sea, el total: C1 = 13 rangos.
Ídem, es C2 = 13
14; 15 O sea, el total es C3 = 2
12; 9; 10; 6,5; 6,5; 8; 14; 11; 15. O sea, el total es C4 = 9
4; 12; 9; 10; 6,5; 6,5; 8; 14; 11; 15. Esto es, C5 = 10
Ídem anterior: C6 = 9
14; 15. O sea, C7 = 2
10; 14; 11; 15. Esto es, C8 = 4
14; 11 ; 15. O sea, C9 = 3
(6,5); 8; 14; 11; 15. En empate se suma medio punto: C10 = 4,5
8; 14; 11; 15. Ahora será: C11 = 4
14; 11; 15. Estoes, C12 = 3
Solo hay una mayor: 15. Por lo tanto, C13 = 1
Ídem: C14 = 1
Y C15 = 0

Paso 2) Se calcula la suma de puntajes Ci como: $C = \sum Ci = 13 + 13 + 2 + ... + 1 = 78,5$

Paso 3) Se obtiene el valor del ordenamiento con:

$$O = 4 \sum Ci - N(N-1) = 4\,C - N(N-1) = 4\,(78{,}5) - [\,15\,(15-1)] = 314 - 210 = 104$$

Paso 4) Se calcula el estadígrafo de Kendall con:

Como hay empates $$\tau = \frac{O}{\sqrt{[N(N-1) - \sum E_2][N(N-1) - \sum E_1]}}$$

Donde: $\sum E_1 = 0$ porque Y1 no tiene empates.

Y $\sum E2 = 2$ porque hay 2 empates en Y2 con los rangos 6 y 7

Entonces:

$$\tau = \frac{104}{\sqrt{[15(15-1) - 0][15(15-1) - 2]}} = 104 / 209 = 0{,}4976$$

Paso 5) para comprobar la hipótesis de que hay correlación se puede usar la Tabla 21 del Anexo, o si las muestras son grandes, se puede usar la aproximación normal.

$\tau = 0{,}4976*$ versus $\tau_{\alpha;N}$ De tablas es $\tau_{0,95;\,15} = 0{,}390$ y $\tau_{0,99;\,15} = 0{,}505$

Luego se han encontrado resultados significativos que prueba que hay correlación: Esto es, se rechaza Ho : $\tau = 0$.

Realmente la tabla 21 es exacta cuando no hay empates. En casos como el presente solo es aproximada y conviene usar una tabla especial presentada por Burr (1960).

22.7 Problemas propuestos

1) Marcar la respuesta correcta a cada una de las afirmaciones siguientes, o completar la frase:

1) El coeficiente de correlación se usa para medir el grado de asociación de 2 variables.	**V**	**F**
2) Las relaciones matemáticas entre los coeficientes b y r son muy estrechas y sin sentido.	**V**	**F**
3) Si X e Y son aleatorias corresponde hacer un Modelo II de regresión.	**V**	**F**
4) La correlación se usa cuando el investigador busca establecer el grado de asociación.	**V**	**F**
5) Presentar en un cuadro los 4 casos posibles sobre que hacer con Regresión y Correlación...	.....	
6) El coeficiente r es el cociente entre covarianza y el producto de los DSx y DSy .	**V**	**F**
7) Explicar la fórmula del producto-momento de Pearson:.....	.....	
8) El cociente entre la VE de X y la VT de Y es el coeficiente de correlación.	**V**	**F**
9) El cuadrado del coeficiente de regresión es el coeficiente de determinación .	**V**	**F**
10) Explicar los pasos a seguir para seguir los cálculos en correlación.	.....	
11) Explicar los pasos para testear si hay correlación:.....	.....	
12) Para testear correlación hay dos métodos: el de Student y otro con la Tabla 20.	**V**	**F**
13) La transformación de Fisher permite usar la Chi cuadrado para el test de correlación.	**V**	**F**

14) El ln[(1+r)/(1-r)] transforma r en Z. **V** **F**
15) Los grados de libertad de Z son N-3. **V** **F**
16) La tipificación de Z permite testear Ho) $\rho = a$ **V** **F**
17) Si las muestras son grandes n > 50 se puede aproximar z con gauss. **V** **F**
18) Si N está entre 10 y 25 se debe efectuar la corrección de Hotelling. **V** **F**
19) Cuando se compara más de dos muestras, se testea que todas provengan de una población **V** **F**
20) Para estimar el valor poblacional de r no se puede usar un promedio directo. **V** **F**
21) Explicar los pasos a seguir en un test de homogeneidad de r en más de 2 muestras:..... .....
22) El modelo de Wilcoxon se puede usar para testear si hay correlación. **V** **F**
23) El modelo no paramétrico para la correlación es el de:..... .....
24)Las magnitudes deben ser por lo menos ordinales en un test no paramétrico de correlación **V** **F**

2) Para los datos de la tabla siguiente decidir si:
a) Existe correlación entre las dos variables.
b) Si $\rho = 0,6$

Nº	X	Y
1	80	5
2	100	15
3	120	25
4	140	44
5	160	67
6	180	72
7	200	81
8	220	90
9	240	106
10	260	120

3) Resolver el mismo problema anterior con el modelo de Kendall.

4) Resolver el siguiente problema con el modelo de Kendall.

N	Y1	Y2
1	10	120
2	11	105
3	14	120
4	22	130
5	25	129
6	29	140
7	33	160
8	38	154
9	41	152
10	45	170
11	52	190
12	55	180