Correlación No-Paramétrica y su Aplicación en la Investigaciones Científica
Non-Parametric Correlation and Its Application in Scientific Research

Badii, M.H., A. Guillen, O.P. Lugo Serrato & J.J. Aguilar Garnica
UANL, San Nicolás, N.L., México, mhbadiiz@gmail.com

Resumen: En este trabajo, se describen y se expliquen las técnicas no paramétricas de correlación y sus aplicaciones a través de ejemplos. Por su relevancia y aplicación específica, se pone un énfasis particular sobre la prueba de rango de Spearman, la prueba de tau de Kendall, la prueba de concordancia de Kendall, la Prueba de Kappa y la prueba bi-serial de puntos.
Palabras clave: Distribución libre, técnicas no paramétricas de correlación,

Abstract: In this essay, non-parametric techniques of correlation are described and explained with examples for each case. Due to their relevance and their specific application, a particular emphasis is placed upon some non-parametric correlation techniques such as Spearman Rank Correlation, Kendall´s Tau Test, Kendall´s Concordance Test, Kappa´s Test, and Bi-serial points Test.
Keywords: Free distribution, non-parametric correlation techniques

Concepto

En las investigaciones, existe un interés en conocer si existe asociaciónón de algún tipo ya sea positiva o negativa entre variables bajo el estudio. La técnica de correlación se encarga a determinar este grado de asociación o correlación (Badii et al, 2007, 2009). Fue Pearson (Pearson, 1929) quien descubrió la ecuación de correlación paramétrica y también fue él que determinó la ecuación del coeficiente de correlación el cual mide el grado de correlación entre dos variables "X" y "Y". El coeficiente de correlación fluctúa de -1 indicando una correlación negativa perfecta de 100%, hasta +1 que a su vez indica una correlación de 100% pero en este caso positiva. Cabe indicar que en el caso de correlación negativa, un aumento en una variable se asocia con un decremento en otra variable, y en el caso a la correlación positiva ocurre contrario a lo mencionado anteriormente. Correlación paramétrica se aplica para casos en donde la distribución de los datos sigue una curva Gausiana o normal. Sin embargo, los datos conseguidos en las ciencias naturales y ciencias sociales, raramente se ajustan a la curva normal, y por esto cuando se tarat de estudiar el grado de correlación en estas ciencias uno debe utilizar técnicas de correlación no-paramétrica, las cuales son libre de distribución, es decir, no existe la necesidad de que los datos tengan una distribución normal. En este trabajo se desciben y discuten con ejemplos en forma detallada y paso por paso el uso y la aplicación de las técnicas de correlación no-paramétricas (Spearman, 1904, kendall, 1938, Zar, 1973).

CORRELACIÓN DE SPEARMAN

Cuando las dos variables bajo del estudio de correlación no tienen distribución normal se procederá con los rangos de mediciones para cada variable. Hay dos métodos de rango de correlación, uno de Spearman (1904) y otro de kendall (Kendall, 1938, Kendall & Babington-Smith, 1939).
En caso de rango de correlación de Spearman, después de dar los rangos a cada medición de la variable se usará la ecuación siguiente para proceder con la operación.

$$r_s = 1 - [6 \sum d_i^2 / (n^3 - n)]$$

Donde, d_i = diferencia entre rangos de X y Y.
El valor de r_s varía de "-1" hasta "+1" y no tiene unidad, sin embargo, este valor es diferente del valor de *r* calculado por el método de Pearson.

Ejemplo. Prueba de Rango de Spearman sin empates.

Los datos siguientes (Tabla 1) son calificaciones de 10 alumnos en dos materias distintas en la universidad (Pérez Tejada, 2008).

Tabla 1. Calificaciones de alumnos en dos materias.

Alumno	Calificación en materia 1 (Xi)	Rango de Xi	Calificación en materia 2 (Yi)	Rango de Yi	Di (Xi – Yi)	di^2
1	57	4	83	7	-4	16
2	45	1	37	1	0	0
3	72	7	41	2	5	25
4	78	8	84	8	0	9
5	53	2	56	3	-1	1
6	63	5	85	9	-4	16
7	86	9	77	6	3	9
8	98	10	87	10	0	0
9	59	4	70	5	-1	1
10	71	6	59	4	2	4

$n = 10$

$$\sum d_i^2 = 72$$

$$r_s = 1 - [6 \sum d_i^2 / (n^3 - n)] = 1 - [6(72) / (10^3 - 10)] = 1 - 0.436 = 0.564$$

Ho: $r_s = 0$

Ha: $r_s \neq 0$

$(r_s)0.05(2), 10 = 0.648$

El valor 0.648 viene de la Tabla de Spearman con $n = 10$ y $\alpha = 0.05$ para prueba bilateral.

Valor calculado (0.564) es menor que valor tabulado (0.648), y por tanto Ho se acepta a nivel de $\alpha = 0.05$, es decir con 95% de confianza se dictamina que no hay correlación entre las calificaciones de dos materias.

Ejemplo. Prueba de Rango de Spearman con empates.
Los datos siguientes (Tabla 2) representan la longitud de ala (Xi) y de cola (Yi) de 12 aves.

Tabla 2. Longitud (cm) de alas y colas en 12 aves.

Xi	Rango de Xi	Yi	Rango de Yi	di	di^2
10.4	4.0	7.4	5.0	-1.0	1.00
10.8	8.5	7.6	7.0	1.5	2.25
11.1	10.0	7.9	11.0	-1.0	1.00
10.2	1.5	7.2	2.5	-1.0	1.00
10.3	3.0	7.4	5.0	-2.0	4.00
10.2	1.5	7.1	1.0	0.5	0.25
10.7	7.0	7.4	5.0	2.0	4.00
10.5	5.0	7.2	2.5	2.5	6.25
10.8	8.5	7.8	9.5	-1.0	1.00
11.2	11.0	7.7	8.0	3.0	9.00
10.6	6.0	7.8	9.5	-3.5	12.25
10.4	12.0	8.3	12.0	0.0	0.00

$n = 12$

$\sum \mathbf{d_i^2 = 42}$

$\mathbf{r_s = 1 - [6 \sum d_i^2 / (n^3 - n)] = 1 - [6(42) / (12^3 - 12)] = 1 - 0.147 = 0.853}$

Ho: $r_s = 0$

Ha: $r_s \neq 0$

$(r_s)_{0.05(2),12} = 0.587$ este valor viene de la Tabla de Spearman con $n = 12$ y $\alpha = 0.05$ para prueba bilateral.

Valor calculado (0.853) es mayor que valor tabulado (0.583), y por tanto Ho se Rechaza a nivel de $\alpha = 0.05$, es decir, con 95% de confianza sí existe una correlación positiva entre las mediciones o valores de Xi y Yi.

COEFICIENTE TAU (T) DE KENDALL

El coeficiente tau (τ) de kendall está basada más en los intervalos jerarquizados de las observaciones que los propios datos, esto hace que la distribución de τ sea independiente de la que presentan las variables X y Y, siempre y cuando que los datos representados por estas 2 variables sean (1) independientes y (2) continuas. Este coeficiente es más preferida por algunos investigadores que el de Spearman, pero es más difícil de calcular, pero con una ventaja de que el τ tiende más rápido a la distribución normal que el de Spearman, especialmente, en el caso de la certeza de Ho.

Ecuación.

$\mathbf{\tau = (Sa - Sb) / [n(n - 1) / 2]}$

Donde,

τ = Estadística de Kendall

n = # de casos en el ejemplo

Sa = Sumatoria de rangos más altos

Sb = Sumatoria de rangos más bajos

Ejemplo. En una evaluación de los jugadores delanteros de futbol en de un país, hay 9 de ellos catalogados como más intensos para marcar goles. Para analizar esta intensidad durante un periodo de una temporada se registro sistemáticamente el grado de intensidad de cada uno

de estos delanteros tanto en juegos a nivel nacional (NP = puntajes nacional), como a nivel internacional (IP = puntajes en juegos internacionales). Además, se registraron los rangos a nivel nacional (NR = rangos a nivel nacional) y en a nivel internacional (IR = rango a nivel internacional). Los datos se presentan en la Tabla 3. Los rangos se ordenan de máxima a mínima hacia abajo en cada columna de rango.

Tabla 3. Datos de grado de agresividad en la guardería y el hogar.

Jugador	NP	IP	NR	IR
1	84	60	1	4
2	80	64	2	2
3	78	71	3	1
4	76	61	4	3
5	70	58	5	5
6	64	57	6	6
7	62	54	7	8
8	50	55	8	7
9	47	52	9	9

Procedimiento.

Paso 1.

Se considera el IR como referencia y comienza a contabilizar a partir del primer rango, es decir, el rango con el valor de 4 y cuenta el número de los rangos menores que 4 (hacia debajo de 4): en este caso los tres números de 2, 1, y 3, es decir tenemos 3 valores menores que el valor 4. Luego cuentan los rangos mayores de 4 a partir e incluyendo el número 5, así tenemos los valores 5, 6, 8, 7, y 9, es decir, hay 5 rangos mayores que el valor 4. Se continúa así contabilizar los rangos menores y mayores para los siguientes valores de la columna de IR, es decir, a partir del valor 2 en adelante. De esta manera se generan los valores de las 2 columnas de Sa (sumatoria de rangos más altos) y Sb (sumatoria de rangos más bajos, Tabla 4).

Tabla 4. Datos de grado de agresividad en la guardería y el hogar y los de Sa y Sb*.

Jugador	NP	IP	NR	IR	Sa = 31	Sb = 5
1	84	60	1	4	5	3
2	80	64	2	2	6	1
3	78	71	3	1	6	0
4	76	61	4	3	5	0
5	70	58	5	5	4	0
6	64	57	6	6	3	0
7	62	54	7	8	1	1
8	50	55	8	7	1	0
9	47	52	9	9	0	0

*: Sa = sumatoria de rangos más altos, Sb = sumatoria de rangos más bajos.

Ahora substituir en la ecuación de Kendall resulta: $\boldsymbol{\tau = (Sa - Sb) / [n(n-1) / 2]} = (31 - 5) / [9(9 - 1)/2] = 26 / 36 = 0.72$, hay una asociación de 72%.

COEFICIENTE DE CONCORDANCIA (ώ) DE KENDALL

Se considera a este coeficiente como un promedio de un grupo de coeficientes de Spearman, es decir, el ώ es una medida del grado de acuerdo (concordancia) entre *m* conjuntos de *n* rangos. Por ejemplo, para un grupo de *n* objetos evaluados por *m* jueces, la ώ provee información sobre el grado de acuerdo entre *m* rangos otorgados por los jueces. Un diferencia entre r_S (coeficiente de Spearman) y el ώ es que el de kendall siempre es un valor positivo entre 0 y 1, es decir, sí la evaluación de cada juez a los *n* objetos es similar, entonces la ώ es igual a 1, en cambio si hay un total desacuerdo, entonces el ώ = 0. Sin embargo, hay que tomar en cuenta que un ώ = 0 puede indicar que los atributos a evaluar son ambiguos o están pobremente definidos, consecuentemente, no se puede discriminar y por tanto hay desconcordancia

Ecuación (Kendall, 1938, Kendall & Babington-Smith, 1939).

$$\dot{\omega} = 12\sum D^2 / m^2 n(n^2 - 1)$$

Para comprobar la validez de los resultados se usa la ecuación: $\sum R = [mn(n + 1)] / 2$.

Donde,

ώ = Estadística de Kendall
$D = \sum R - (\sum R/n)$
m = Rango de evaluadores
n = Número de objetos
$\sum R$ = Suma de rangos

Ejemplo.

Se ofrece 5 diferentes técnicas de explotar y extraer el gas como fuente de energía de la tierra a 4 expertos en la materia ($n = 5$, $m = 4$). Los expertos ranquean la calidad de cada técnica en función del daño posible a la salud humana y al medio ambiente de 1 (más bajo) a 5 (más alto) según los datos en la Tabla 5.

Tabla 5. Los valores de los rangos de los expertos sobre la calidad de explotar y extraer el gas.

Rango	Experto I	Experto II	Experto III	Experto IV	Suma de rangos
1	5	4	5	5	19
2	3	3	2	3	11
3	1	2	1	2	6
4	2	1	3	1	7
5	4	5	4	4	17
					$\sum R = 60$

Procedimiento.

1. Sumar los rangos por cada objeto, es decir, arrojar finalmente, $\sum R = 60$.

2. Si $\dot{\omega} = 0$, es decir total desconcordancia, sería como una evaluación aleatoria, y por tanto tendremos que $\sum R / n = 60 / 5 = 12$, este valor (12) sería el promedio de rangos en el caso de $\dot{\omega} = 0$.

3. Por tanto la diferencia (D) sería igual a $D = \sum R - 12$.

4. Construir la Tabla 6 que contenga las columnas de la Tabla anterior más la D.

5. Substituir en la ecuación.

Tabla 6. Los rangos de los expertos sobre la calidad de explotar y extraer el gas*.

Rango	**Exp I**	**Exp II**	**Exp III**	**Exp IV**	**Suma de rangos (∑R)**	**|D| = ∑R – 12**	**D^2**
1	5	4	5	5	19	7	49
2	3	3	2	3	11	1	1
3	1	2	1	2	6	6	36
4	2	1	3	1	7	5	25
5	4	5	4	4	17	5	25
					∑ total = 60		∑=136

*: Exp = Experto, | = Tomar el valor absoluto

$\dot{\omega} = 12\sum D^2 / m^2 n(n^2 - 1) = 12(136) / 4^2\ 5(5^2 - 1) = 1.632 / 1{,}920 = 0.85$

Comprobación:

$\sum R = [mn(n + 1)] / 2$

$60 = 4*5*6 / 2$

$60 = 120 / 2 = 60$

COEFICIENTE DE CORRELACIÓN (R_{BS}) BI-SERIAL DE PUNTOS

Este coeficiente es una medida de asociación entre 2 variables continuas, en donde una de ellas es dicotómica. La variable dicotomizada se supone discreta o discontinua cuando trata de relacionarse con la que permanece continua.

Ecuación

$$\mathbf{r_{bs} = n(\sum f_c X) - n_c(\sum fX) / \{(n_c)(f_i)\ [n(\sum (fX)^2 - (\sum fX)^2]\}^{1/2}}$$

Donde,

r_{bs} = estadística o coeficiente bi-serial
n = # de casos
f = frecuencia de las puntuaciones obtenidas por los sujetos
X = puntuación obtenidas por los sujetos
f_c = # de sujetos que obtuvieron exactamente las puntuaciones X
n_c = # total de sujetos que obtuvieron exactamente las puntuaciones X
f_i = # de sujetos que NO obtuvieron exactamente las puntuaciones X
n_i = # total de sujetos que NO obtuvieron exactamente las puntuaciones X

Ejemplo

A un total de 100 alumnos (Pérez Tejada, 2008) se aplica una prueba de conocimiento X que consta de 40 preguntas. Los resultados se encuentran en la Tabla 7.

Tabla 7. Resultados del examen de 40 preguntas (X) y sus frecuencias (f) (n = 100).

X	**f**
40	2
38	4
37	6
36	12
32	12
31	10
30	12
28	10
27	10
25	4
24	4
22	3
20	3
18	3
16	2
12	2
10	1

Durante la realización del examen se enfatizó la pregunta # 23, es decir, el criterio para evaluar era correcto o incorrecto. Se calculó el coeficiente r_{bp} vía puntuaciones obtenidas por 100 alumnos en la prueba X y también la respuesta a la pregunta # 23 lo cual se dicotomizó como "incorrecto-correcto."

Procedimiento

Se calculan los valores de fc y fi (Tabla 8).

Tabla 8. Resultados del examen de 40 preguntas y sus frecuencias*.

X	f	Fc	fi	fX	$f(X^2)$	fcX
40	2	2	0	80	3200	80
38	4	4	0	152	5776	152
37	6	5	1	222	8214	185
36	12	10	2	432	15552	360
32	12	9	3	384	12288	288
31	10	8	2	310	9610	248
30	12	7	5	360	10800	210
28	10	6	4	280	7840	168
27	10	7	3	270	7290	189
25	4	1	3	100	2500	25
24	4	1	3	96	2304	24
22	3	1	2	66	1452	22
20	3	1	2	60	1200	20
18	3	0	3	54	972	0
16	2	1	1	32	512	16
12	2	0	2	24	288	0
10	1	0	1	10	288	0
Totales	**100**	**63**	**37**	**2932**	**89898**	**1987**
	$\sum f = n$	nc	ni	$\sum fX$	$\sum fX^2$	$\sum fcX$

*: fc = Frecuencia de los alumnos con respuesta correcta a la pregunta # 23. fi = Diferencia entre la columna f o total de alumnos para cada pregunta & la columna fc; es decir, frecuencia de alumnos que NO respondieron correctamente la pregunta # 23.

$$r_{bs} = n(\sum f_c X) - n_c(\sum fX) / \{(n_c)(n_i)\ [n(\sum(f(X^2) - (\sum f_x)^2]\}^{1/2}$$

$r_{bs} = [(100)(1987) - 63(2932)] / \{(63)(37)\ [100\ (89898) - (2932)^2]\}^{1/2} = 13984/30274.641 = 0.4619$

PRUEBA DE KAPPA

$K = (P_0 - Pe) / (1 - Pe)$

El valor de K varía de 0 a 1, donde, 1 es confiabilidad perfecta y 0 = no hay confiabilidad
Ahora, valores siguientes se consideran de para dictaminar diferentes niveles de confiabilidad:
K > 0.75: confiabilidad excelente
K entre 0.75 y 0.4: confiabilidad buena.
K < 0.4: confiabilidad pobre.

Descripción

Medir la confiabilidad de mediciones hechas por el mismo experto o dos expertos (de forma independiente) cuando la variable es categórica y subjetiva (dolor, inflamación, efectividad del medicamento, etc.), por tanto se necesita establecer una escala por ejemplo: 0 = no hay, 1 = leve, 2 = regular, 3 = severo.

Ejemplo

Evaluación de tumores de 90 pacientes por 2 médicos considerando la siguiente escala: 0 = No tumor, 1 = tumor de grado leve, 2 = tumor de grado regular, 3 = tumor de grado severa.

Se ubican las evaluaciones de 2 médicos en una tabla de contingencia de 2 x 2 (Tabla 9).

Tabla 9. Niveles subjetivas de tumores indicados por 2 médicos.

		Médico "II"				**Total**
	Nivel del tumor	**0**	**1**	**2**	**3**	
Médico "I"	**0**	**25**	2	1	0	28
	1	4	**17**	3	2	26
	2	1	3	**15**	1	20
	3	0	1	2	**13**	16
Totales		30	23	21	16	GT = 90

$K = (P_0 - Pe) / (1 - Pe)$, Donde,

Po = Proporción observada de respuesta similares para ambos expertos
Pe = Proporción esperada de respuesta similares para ambos expertos
Pe = $\sum$ QiBi
Qi = Proporción de respuesta en categoría "i" para el experto I
Bi = Proporción de respuesta en categoría "i" para el experto II
Po = (25 + 17 + 15 + 13) / 90 = 0.78
Po = Suma de valores de diagonal de la Tabla 9
Valor diagonal = Respuesta de 2 expertos al nivel misma de la escala

Q0 = 28/90 =0.31
Q1 = 26/90 =0.29
Q2 = 20/90 =0.22
Q3 = 16/90 =0.18

B0 = 30/90 = 0.33
B1 = 23/90 = 0.26
B2 = 21/90 = 0.23
B3 = 16/90 = 0.18
Pe = $\sum$ QiBi = (0.31 * 0.33) + (0.29 * 0.26) + (0.22 * 0.23) + (0.18 * 0.18) = 0.26
$K = (P_0 - Pe) / (1 - Pe)$ = (0.78 - 0.26) / (1 - 0.26) = 0.70
Por tanto, hay una confiabilidad buena en este caso.

Conclusión

La meta final de una investigación científica es 1) el poder distinguir entre grupos de datos, es decir, el establecer si los grupos distintos de datos se originan de diferentes universos o poblaciones diferentes o en realidad pertenecen a la misma población y por tanto, son

segmentos del mismo universo, y 2) ser capaz de detectar la existencia o no de patrones o hechos repetitivos en espacio y tiempo con la misma función o comportamiento.

En la mayoría de las situaciones los investigadores tratan de utilizar datos que tiene distribución normal, sin embargo, la vida y casi todo relacionado con aspectos de las ciencias sociales y humanistas y naturales no poseen datos con distribución normal y de hecho, la distribución predominante para estos asuntos tiende a log-normal. La técnica de correlación junto con regresión apoya a distinguir los patrones. En este trabajo se toman en cuenta aquellas técnicas de correlación con distribución libre o no-paramétrico y demuestra sus aplicaciones en la investigación científica.

Referencias

Badii, M.H., J. Castillo, J. Rositas & G. Alarcón. 2007. Uso de un método de pronóstico en investigación. Pp. 137-155. In: M.H. Badii & J. Castillo (eds.). Técnicas Cuantitativas en la Investigación. UANL, Monterrey.

Baddi, M.H., J. Castillo, J. Landeros & K. Cortez. 2009. Papel de la estadística en la investigación científica. Pp. 1-43. In: M.H. Badii & J.Castillo (eds). Desarrollo Sustentable: Métodos, Aplicaciones y Perspectivas. UANL. Monterrey.

Kendall, M.G. 1938. A new measure of rank correlation. Biometrika, 30: 81-93.

Kendall, M.G. & B. Babington-Smith. 1939. The problem of *m* rankings. Ann. Math. Statist. 18: 495-513. 1939.

Pearson, K. 1920. Notes on the history of correlation. Biometrika, 25-45.

Pérez Tejada, H.E. 2008. Estadística para las Ciencias Sociales, del Comportamiento y de la Salud. Cengae Learninig. Auatralia.

Spearman, C. 1904. The proof and measurement of association between two things. Am. J. Psychol. 15: 72-101.

Zar, J.H. 1973. Significance testing of the Spearman Rank Correlation Coefficient. J. Amer. Statist. Assoc. 67: 578-580.