

CORRELACIÓN POLICÓRICA EN EL ANÁLISIS DE FACTORES CON VARIABLES ORDINALES

Ramos Barberán Miriam¹, Plata Alarcón Wendy²

Resumen En este artículo se presenta la explicación del cálculo de las correlaciones policóricas a partir de variables ordinales. Para el efecto, se construyen variables latentes a partir de las ya observadas, estas variables definen umbrales en el dominio de distribuciones normales, con las cuales se calculan las correlaciones. Bajo estas condiciones, se ilustra una aplicación del uso de las correlaciones policóricas en el análisis de factores exploratorio para identificar los grupos de variables que influyen en la satisfacción al cliente del Departamento de Postgrado en una universidad.

Palabras claves: variables ordinales, análisis de factores, correlaciones tetracóricas, correlaciones policóricas, máxima verosimilitud.

Abstract. This paper shows an explanation of the polychoric correlation calculus from ordinal data. For this purpose, latent variables are created from observed ones, these variables define thresholds in the normal distribution domain, which are used to calculate the correlations. Based on those conditions, we shall illustrate the polychoric correlation applied to exploratory factor analysis for identifying a set of variables that affect customer satisfaction in the Postgraduate Department of a university.

Keywords: ordinal variable, factors analysis, tetrachoric correlation, polychoric correlation, maximum likelihood.

Recibido: Enero 2015.

Aceptado: Febrero 2015.

1. INTRODUCCIÓN

El uso de las variables ordinales se ha generalizado ampliamente en las ciencias sociales, específicamente en la medición de la percepción de los informantes respecto a determinada realidad, tal es el caso de la Psicología al estudiar el comportamiento humano de manera cualitativa, así como también, la Investigación de Mercado al aplicar cuestionarios para medir el grado de satisfacción de los clientes. En el presente artículo, se explora la utilización de la correlación policórica en el análisis de factores con datos ordinales considerando correlaciones de variables subyacentes a las observadas, con las que se espera obtener una mayor explicación de la varianza de los datos.

2. METODOLOGÍA

En este trabajo se inició con la revisión de los métodos utilizados en el análisis de factores a partir de datos ordinales, lo cual condujo a descubrir la importancia de una adecuada estimación de la correlación en este tipo de estudios. En este sentido, se tomó como punto de partida la correlación tetracórica propuesta por Pearson en 1901 para variables dicotómicas; culminando con la respectiva generalización para variables politómicas, que es la denominada correlación policórica. Se presenta la explicación de un método para la estimación del coeficiente de correlación policórica, mediante la maximización de la función de verosimilitud de la muestra, donde

en primera instancia se estimaron los umbrales de las variables latentes que siguen una distribución normal estándar y luego se calculó la correlación.

Finalmente, con los datos del Estudio Transversal de Satisfacción al Cliente del Departamento de Postgrado de una universidad, realizado por Plata [1], se calcularon las correspondientes correlaciones de Pearson y Policóricas para identificar diferencias entre ellas; y, se comparó el porcentaje de la variabilidad explicada en el análisis de factores.

3. VARIABLES ORDINALES Y ANÁLISIS DE FACTORES

El uso de variables ordinales es muy común para medir percepción respecto a diferentes circunstancias de ámbito social, siendo así que pueden definirse categorías con la finalidad de graduar la calificación del informante respecto a su satisfacción frente a determinada característica. Para el efecto, se estila usar escalas que asignan valores enteros a cada categoría, tales como la escala de Likert, véase Tabla I.

TABLA I:

Se estila usar escalas que asignan valores enteros a cada categoría, tales como la escala de Likert

Categoría de Variables Ordinales

Nada satisfecho	Poco satisfecho	Indiferente	Satisfecho	Muy satisfecho
1	2	3	4	5

Bajo este mismo esquema se puede definir un gran conjunto de variables a ser analizadas de forma univariada y multivariada, donde para el

¹Ramos Miriam MPC Profesora Departamento Matemáticas, Facultad Ciencias Naturales y Matemáticas ESPOL (e-mail: mvramosb@espol.edu.ec), Guayaquil.

²Plata Wendy MPC Profesora Departamento Matemáticas, Facultad Ciencias Naturales y Matemáticas ESPOL (e-mail: wplata@espol.edu.ec), Guayaquil.

caso último la técnica estadística frecuentemente usada es el análisis factorial cuyo objetivo es reducir la redundancia entre las variables empleando un pequeño número de factores [2]. Olsson [3] mostró que la aplicación del análisis de factores con datos discretos puede conducir a inferencias erróneas respecto al número de factores y a estimaciones sesgadas de las cargas de los factores, lo cual se atribuye a la estimación sesgada de las correlaciones.

4. ESTIMACIÓN DEL COEFICIENTE DE CORRELACIÓN POLICÓRICA

Con el propósito de resolver el problema identificado por Olsson respecto a la estimación sesgada de las correlaciones en el análisis de factores, se introduce el estudio de la correlación policórica que parte del supuesto que existen variables no observables (latentes) de naturaleza continua, normalmente distribuidas, sobre las cuales se han construido variables ordinales observables, cuyas opciones de respuesta se corresponden con umbrales en el dominio continuo de las latentes, este tipo de correlación es usada en análisis de factores que involucre variables puramente ordinales según Richaud [4].

Sean x e y dos variables ordinales cuyas opciones de respuesta (categorías) están dadas por r y s respectivamente; y, sean ξ y η las correspondientes variables latentes normales bivariadas, bajo estas condiciones, los umbrales en el dominio continuo se definen como sigue, Olsson [5]:

Para x :

$$\begin{aligned} x = 1 & \text{ si } \xi < a_1 \\ x = 2 & \text{ si } a_1 \leq \xi < a_2 \\ x = 3 & \text{ si } a_2 \leq \xi < a_3 \\ x = 4 & \text{ si } a_3 \leq \xi < a_4 \\ & \vdots \\ x = s & \text{ si } a_{s-1} \leq \xi \end{aligned}$$

Para y :

$$\begin{aligned} y = 1 & \text{ si } \eta < b_1 \\ y = 2 & \text{ si } b_1 \leq \eta < b_2 \\ y = 3 & \text{ si } b_2 \leq \eta < b_3 \\ y = 4 & \text{ si } b_3 \leq \eta < b_4 \\ & \vdots \\ y = r & \text{ si } b_{r-1} \leq \eta \end{aligned}$$

Siendo a_i y b_j los umbrales de las variables ordinales x e y , respectivamente, tal que:

$$a_0 = b_0 = -\infty \quad a_s = b_r = +\infty$$

Con las variables ordinales se construye una Tabla de Contingencia, que consiste en un arreglo rectangular de $s \times r$ celdas, donde cada una de ellas representa las frecuencias observadas para la distribución conjunta de las variables x e y , a partir de las cuales subyacen las variables latentes ξ y η con sus respectivos umbrales a_i y b_j , véase TABLA II.

TABLA II:
Consiste en un arreglo rectangular de $s \times r$ celdas
Tabla de Contingencia

x \ y	y						
	1	2	3	4	...	b_{r-1}	r
a ₁	n ₁₁	n ₁₂	n ₁₃	n ₁₄			n _{1r}
	n ₂₁	n ₂₂	n ₂₃	n ₂₄			n _{2r}
	n ₃₁	n ₃₂	n ₃₃	n ₃₄			n _{3r}
	n ₄₁	n ₄₂	n ₄₃	n ₄₄			n _{4r}
a _{s-1}							
	n _{s1}	n _{s2}	n _{s3}	n _{s4}			n _{sr}

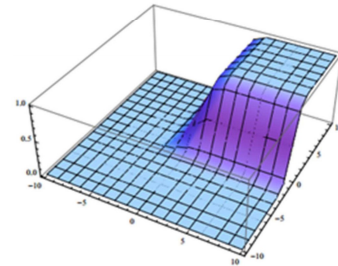
El problema a resolver consiste en calcular la correlación policórica ρ como una medida de la relación lineal entre las variables latentes ξ y η . Esta correlación constituye una generalización de la correlación tetracórica introducida por Pearson en 1901 [6], para tablas de contingencia de orden 2×2 . En este sentido, la correlación policórica ρ es una medida que se obtiene a partir de la distribución acumulada normal bivariada:

Exp. (1)

$$\Phi(h, k; \rho) = \left[2\pi(1 - \rho^2)^{\frac{1}{2}} \right]^{-1} \int_{-\infty}^h \int_{-\infty}^k e^{\left[\frac{x^2 - 2\rho xy + y^2}{2(1 - \rho^2)} \right]} dx dy$$

A manera de ilustración de Exp. (1) se presenta la FIGURA I.

FIGURA I:
Distribución Acumulada
Normal Bivariada $\Phi(10,10; 0.5)$



Según Zurita [7], a la distribución conjunta de una muestra se la considera como una función de parámetros de la población, denominada Función de Verosimilitud de la Muestra, la cual en general se utiliza para la estimación de los parámetros que identifican a una población. Considerando las frecuencias ilustradas en la TABLA II y teniendo como propósito la estimación de la correlación policórica ρ y de los umbrales a_i y b_j , la función de verosimilitud de la muestra para este caso particular está dada por:

Exp. (2)

$$L = C \cdot \prod_{i=1}^s \prod_{j=1}^r \pi_{ij}^{n_{ij}}; i = 1, 2, 3, 4, \dots, s; j = 1, 2, 3, 4, \dots, r$$

$$L = C \cdot (\pi_{11}^{n_{11}})(\pi_{12}^{n_{12}})(\pi_{13}^{n_{13}})(\pi_{14}^{n_{14}}) \dots (\pi_{sr}^{n_{sr}})$$

Donde, C es una constante, n_{ij} representa la frecuencia observada en la celda ubicada en la i -ésima fila y j -ésima columna; mientras que, π_{ij} es la probabilidad de que una observación caiga en la antes referida celda, tal que:

Exp. (3)

$$\pi_{ij} = \Phi_2(a_i, b_j) - \Phi_2(a_{i-1}, b_j) - \Phi_2(a_i, b_{j-1}) + \Phi_2(a_{i-1}, b_{j-1})$$

En Exp. (3) Φ_2 corresponde a la distribución acumulada normal bivariada con correlación ρ . A fin de facilitar la manipulación de las expresiones matemáticas, comúnmente se aplica el logaritmo natural a ambos miembros en Exp. (2), con lo cual se obtiene:

Exp. (4)

$$l = \ln L = \ln C + \ln[(\pi_{11}^{n_{11}})(\pi_{12}^{n_{12}})(\pi_{13}^{n_{13}}) \dots (\pi_{sr}^{n_{sr}})]$$

$$l = \ln C + n_{11} \ln(\pi_{11}) + n_{12} \ln(\pi_{12}) + n_{13} \ln(\pi_{13}) + \dots + n_{sr} \ln(\pi_{sr})$$

$$l = \ln C + \sum_{i=1}^s \sum_{j=1}^r n_{ij} \ln(\pi_{ij})$$

Siendo Exp.(4) la expresión generalizada que resulta de la aplicación de logaritmos a la función de verosimilitud de la muestra.

Con el fin de determinar los valores de los parámetros ρ , a_i y b_j , que maximizan la función de verosimilitud, se procede a derivar parcialmente Exp. (4) con respecto a ρ y a igualar esta derivada a cero.

Exp. (5)

$$\frac{\partial l}{\partial \rho} = 0$$

$$\sum_{i=1}^s \sum_{j=1}^r \frac{n_{ij}}{\pi_{ij}} [\phi_2(a_i, b_j) - \phi_2(a_{i-1}, b_j) - \phi_2(a_i, b_{j-1}) + \phi_2(a_{i-1}, b_{j-1})] = 0$$

La expresión (5) depende de la función de densidad normal bivariada con correlación ρ , que a su vez proviene de la derivada de la distribución acumulada Φ_2 .

Debido a la complejidad de los cálculos al estimar simultáneamente todos los parámetros, se recurre a un método en cual en primera instancia se estiman los umbrales a_i y b_j , a partir de las marginales proporcionadas por la Tabla de Contingencia (TABLA III), donde $n = n_{11} + n_{12} + \dots + n_{sr}$; $P_{i\cdot}$ corresponde a la suma de las frecuencias de la i -ésima fila dividida para n ; y, $P_{\cdot j}$ es la suma de las frecuencias de la j -ésima columna dividida para n . Los umbrales a_i y b_j representan los percentiles de la Distribución Acumulada Normal Estándar, Martinson, E. & Hamdan, M. [8].

Exp. (6)

$$a_i = \Phi_1^{-1}(P_{i\cdot})$$

Exp. (7)

$$b_j = \Phi_1^{-1}(P_{\cdot j})$$

TABLA III:

Tabla de Contingencia – Marginales

		1	b_1	2	b_2	3	b_3	4	b_4	$\dots$	b_{r-1}	r
a_1	1	n_{11}	n_{12}	n_{13}	n_{14}						n_{1r}	$P_{1\cdot}$
a_2	2	n_{21}	n_{22}	n_{23}	n_{24}						n_{2r}	$P_{2\cdot}$
a_3	3	n_{31}	n_{32}	n_{33}	n_{34}						n_{3r}	$P_{3\cdot}$
a_4	4	n_{41}	n_{42}	n_{43}	n_{44}						n_{4r}	$P_{4\cdot}$
$\vdots$												
a_{i-1}												
s	s	n_{s1}	n_{s2}	n_{s3}	n_{s4}						n_{sr}	$P_{\cdot s}$
		$P_{\cdot 1}$	$P_{\cdot 2}$	$P_{\cdot 3}$	$P_{\cdot 4}$						$P_{\cdot r}$	

Una vez estimados los umbrales, para calcular el coeficiente de correlación policórica ρ , se resuelve el sistema de ecuaciones generado por Exp. (5) a través de métodos numéricos que se implementan en software computacionales.

5. CORRELACIÓN DE PEARSON VS. CORRELACIÓN POLICÓRICA

El análisis de factores tiene como insumo la matriz de correlaciones de las variables a ser investigadas. Tal como se explicó en la sección 3 del presente artículo, al trabajar con variables ordinales pueden surgir errores en la estimación de los factores y sus cargas, por tal motivo, en este apartado se utiliza la matriz de correlación policórica en el análisis factorial de variables ordinales diseñadas para medir la satisfacción al cliente en el Departamento de Postgrado de una universidad del Ecuador. Para el efecto se definió una escala de Likert de 1 a 5, donde el cliente debió ordenar su respuesta respecto a su grado de satisfacción en cuanto a diferentes aspectos o dimensiones, véase TABLA IV.

TABLA IV:

Aspectos, variables y factores

Aspectos/Dimensiones	Variables ordinales	Factores a extraer
Responsabilidad de la alta dirección	14	3
Gestión de Recursos	12	3
Prestación de Servicios – Aspectos resaltados por Profesores	8	3
Prestación de Servicios – Aspectos resaltados por Estudiantes	8	3
Análisis y Mejora	5	1

Se aplicó un análisis de factores exploratorio a las dimensiones mostradas en la TABLA IV, con la finalidad de identificar un número reducido de variables que caractericen a cada uno de los aspectos considerados; con este fin, se tomó como punto de partida la matriz de correlación policórica.

La antes referida matriz fue calculada en el software estadístico R usando el comando “hetcor” de la librería “polycor”; y, paralelamente se realizó el análisis de factores usando la matriz de correlación de Pearson con el objetivo de evidenciar diferencias entre los porcentajes de varianza explicada para cada dimensión, véase TABLA V.

TABLA V:

Porcentaje de Varianza Explicada

Dimensiones	Correlación Policórica	Correlación de Pearson
Responsabilidad con la Dirección	68,90%	64,20%
Gestión de Recursos	76,20%	71,50%
Prestación del Servicio - Aspectos resaltados por Profesores	73,70%	69,70%
Prestación del Servicio - Aspectos resaltados por Estudiantes	76,70%	73,60%
Análisis y Mejora	64,40%	60,30%

De la TABLA V se deduce que las correlaciones policóricas difieren de las de Pearson en cuanto al porcentaje de varianza explicada en el análisis factorial, esto se debe a que las correlaciones policóricas son más altas por ser calculadas en base a variables latentes normales bivariadas.

Para ilustrar esta última afirmación se muestran las comparaciones entre correlaciones de cuatro variables correspondientes a la dimensión Relación con la Alta Dirección. Véase TABLA VI.

TABLA VI:

Correlaciones Policóricas y de Pearson

CORRELACIONES POLICÓRICAS				
Variables	Marco Legal	Malla Curricular	Duración del Programa	Contenido de Materias
Marco Legal	1			
Malla Curricular	0,487	1		
Duración del Programa	0,269	0,621	1	
Contenido de Materias	0,396	0,735	0,594	1
CORRELACIONES DE PEARSON				
Variables	Marco Legal	Malla Curricular	Duración del Programa	Contenido de Materias
Marco Legal	1			
Malla Curricular	0,438	1		
Duración del Programa	0,247	0,567	1	
Contenido de Materias	0,340	0,673	0,544	1

Para todos los casos de la TABLA VI la correlación policórica resultó mayor a la de Pearson. Como se observa las diferencias no son significativamente grandes para los datos analizados; sin embargo, la aplicación de la correlación policórica en el análisis de factores con variables ordinales evidenció ganancias en el porcentaje de explicación de la varianza, lo cual resulta beneficioso para este tipo de estudios.

6. DISCUSIÓN

Tal como sostiene Olsson [3], en el análisis de factores se puede cometer errores en la estimación de las correlaciones, siendo uno de ellos la subestimación de dichas correlaciones, como se evidenció con el objeto de estudio práctico de este artículo, cuando al calcular las correlaciones de Pearson, éstas resultaron menores que las correlaciones policóricas, las mismas que en el análisis de factores afectaron positivamente en el incremento de la explicación de la variabilidad de los datos.

Aun cuando el método de estimación de máxima verosimilitud analizado en este trabajo demanda exigentes cálculos matemáticos, el uso de un lenguaje de programación como R reduce significativamente el tiempo y complejidad de las estimaciones propuestas por Martinson, E. & Hamdan, M. [8].

7. CONCLUSIONES

Al trabajar con datos que involucren variables que responden a un tipo de escala ordinal, tal como la escala Likert, resulta esencial en el análisis factorial el uso de correlaciones policóricas a fin de obtener el máximo porcentaje de explicación de la varianza.

A nivel de resultados se evidenció que el cálculo de las correlaciones policóricas sugiere, en general, mejores estimaciones que las de Pearson.

La intención final de este artículo ha sido promover entre los investigadores interesados en estudiar variables de índole psicométrico la incorporación del uso de las correlaciones policóricas en sus metodología de análisis estadístico de datos con la finalidad de robustecer las estimaciones efectuadas.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

- [1] **PLATA, W. (2014).** “Diseño de un Sistema de Gestión de la Calidad para un Departamento de Postgrado”, ESPOL, Guayaquil-Ecuador.
- [2] **RENCER, A. (2002).** “Methods of Multivariate Analysis”, Segunda Edición, John Wiley & Sons Inc., USA.
- [3] **OLSSON, U. (1979).** “On the robustness of factor analysis against crude classification of the observations”. Multivariate Behavioral Research, 14, 485-500.
- [4] **RICHAUD, M.C. (2005).** “Desarrollos del análisis factorial para el estudio de ítem dicotómicos y ordinales”. Revista Interdisciplinaria, 22(2), 237-251.
- [5] **OLSSON, U. (1979).** “Maximum likelihood estimation of the polychoric correlation coefficient”. Psychometrika, 44, 443-460.
- [6] **PEARSON, K. (1901).** “Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable”. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 195, 1–47.
- [7] **ZURITA, G. (2008).** “Probabilidad y Estadística: Fundamentos y Aplicaciones”, Primera Edición, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador.
- [8] **MARTINSON, E. O. & HAMDAN, M. A. (1975).** “Calculation of the polychoric estimate of correlation in contingency tables”. Applied Statistics, 24, 272-278.