PESQUIMAT, Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos Vol.
II - N°1 Pgs. 57-68 Lima – Perú Ag. 1999

# ANÁLISIS DE DEPENDENCIA POR TRAYECTORIAS EN TABLAS DE CONTINGENCIA BIDIMENSIONALES

*Doris Gómez Ticerán*[*]

## RESUMEN

Se presenta una propuesta del Análisis de Trayectoria con dos variables categóricas, basada en los Coeficientes de Dependencia y Codependencia (Cordeiro, 1990).

En una tabla de contingencia bidimensional donde una de las dimensiones son las especies de algún animal y la otra dimensión son los ambientes que buscan dichos animales para su alimentación, aplicándose las técnicas convencionales se puede estudiar la asociación entre las dos categorías (especies y ambientes) y habiéndose encontrado la asociación , se ajusta algún modelo explicativo de tal situación. Amplia literatura al respecto se encuentra en Fienberg(1980), Agresti(1982), Andersen(1980) entre otros.

Pocas metodologías permiten la verificación de algún tipo de asociación entre las filas o entre las columnas de la tabla (Goodman, 1975), como por ejemplo, entre las especies del problema propuesto desde el punto de vista de sus preferencias por ambientes. Este es un problema de aplicación muy importante en las áreas biológicas como por ejemplo en ecología.

Por otro lado, resulta de mucha mayor importancia para investigadores de dichas áreas, verificar si las especies se atraen (o se repelen) por sí mismas o a través de relaciones entre los grupos de especies. O sea, es de interés estudiar si hay evidencia de asociación entre los **niveles** de las categorías y la posibilidad de ajustarse un modelo explicativo para un nivel de una de las categorías a través de los otros niveles de la misma variable categórica.

Además, en muchas situaciones prácticas puede ser necesario proponer relaciones causales entre los niveles de una categoría como si fuesen los efectos de los otros niveles de la misma variable categórica y éste problema aún no ha sido abordado.

[*] Facultad de Ciencias Matemáticas de la UNMSM. E-mail: yakov@net.telematic.com.pe

En éste contexto y para solucionar problemas de aplicación como el propuesto se ha desarrollado una propuesta teórica cuyo resumen se presenta en el presente artículo. Se propone utilizar los parámetros de dependencia y codependencia (Cordeiro, 1990), para transponer o emular la teoría de Análisis de Trayectoria para variables cuantitativas (Wright, 1934; Li, 1975; Achcar, 1976; Wermuth, 1980) a las tablas de contingencia de doble entrada (Gómez, 1997).

Con el fin de ilustrar los resultados teóricos del presente artículo, se tiene una aplicación usando "Drosophylas" (moscas) y "Leveduras" (alimentos) para verificar qué especies y por qué prefieren ciertas leveduras, es decir, las prefieren por atracción directa entre ellas o por atracción indirecta a través de otras especies.

## MÉTODO

### Transposición de la Metodología de Análisis de Trayectoria en Variables Cuantitativas para Tablas de Contingencia Bidimensional

#### Introducción

Se presenta una propuesta sobre los principales aspectos metodológicos del Análisis de Trayectoria para Datos Categóricos en Tablas de Contingencia Bidimensionales, que permite la realización de los mismos análisis realizados en análisis de trayectoria con variables aleatorias numéricas.

Se sabe que en tablas de contingencia los datos son frecuencias y para variables aleatorias cualitativas ordenadas o no, no se puede hablar de correlaciones entre ellas. Las medidas de asociación más importantes, tratan de analizar las relaciones entre las variables concurrentes y no entre sus niveles.

Muchas veces en aplicaciones prácticas se pretende estudiar las evidencias de asociación entre los niveles de las categorías de las variables cualitativas o la posibilidad de ajustar un modelo explicativo para un nivel a través de los otros niveles de una de las variables cualitativas concurrentes. También, a veces es necesario proponer relaciones

causales entre los niveles de una categoría como si fuesen los efectos de otros niveles de la misma categoría.

Para solucionar los problemas descritos en el párrafo anterior, se propone utilizar los parámetros de dependencia- codependencia de Cordeiro (1990) para transponer o emular el desarrollo del Análisis por Caminos Wright (1918); Wermuth (1980); Wermuth et. al (1983), en tablas de contingencia bidimensionales .

Si el propósito es hacer inferencias, será necesario atribuir modelos probabilísticos a los datos. En ese sentido, los principales modelos probabilísticos que pueden asociarse a las tablas de contingencia son: Producto de Distribuciones de Poisson, Distribución multinomial y Producto de Distribuciones Multinomiales, (Bishop et. al, 1975; Fienberg, 1980) entre otros. Cabe aquí mencionar que los estimadores máximo verosímiles de las frecuencias esperadas en las celdas de la tabla de contingencia, son los mismos, cualquiera sea el modelo probabilístico adoptado (Agresti, 1984). Esto se debe al hecho de que los estimadores para un modelo determinado satisfazen las restricciones impuestas por el otro modelo.
Los aspectos técnicos sobre los modelos probabilísticos para datos categóricos y sus condicionamientos se encuentran en Andersen (1993), Bishop *et al.* (1984), Fienberg (1980), Agresti (1984) entre otros.

**Análisis de Dependencia**

Sean **A** y **B** dos características de clasificación de indivíduos de una población (supuesta infinita), con $A_1, A_2,....,A_I$ y $B_1, B_2, ...,B_J$ los niveles de **A** y **B** respectivamente. Se define:

$$\mathbf{P} = \left( p_{ij} ;\ i = 1,...,I;\quad j = 1,..,J \right)$$

la matriz de probabilidades, donde

$$p_{ij} = P\left(A_i B_j\right)$$

es la probabilidad de que un individuo escogido al azar de ésta población pertenezca a los niveles "i y j" de las características A y B respectivamente.

Suponiendo que $p_{ij} \geq 0$, para todo (i, j) y haciendo:

$$p_{i\bullet} = \sum_{j=1}^{J} p_{ij} \quad \text{para} \quad 1 \leq i \leq J$$

$$p_{\bullet j} = \sum_{i=1}^{I} p_{ij} \quad \text{para} \quad 1 \leq j \leq J$$

entonces, los vectores fila:

$$p_{r,j} = \left(p_{i1},\ldots,p_{iJ}\right) \Big/ p_{i\bullet} \qquad 1 \leq i \leq I$$

y los vectores columna

$$p_{c,ij} = \left(p_{1j},\ldots,p_{Ij}\right) \Big/ p_{\bullet j} \qquad 1 \leq j \leq J$$

son distribuciones de probabilidad.

También son distribuciones de probabilidad, el vector fila marginal

$$p'_{r,\pm} = \left(p_{\bullet 1},\ldots,p_{\bullet J}\right)$$

y el vector columna marginal

$p'_{c,\pm} = \left(p_{1\bullet},\ldots,p_{I\bullet}\right)$ y generalmente son denominados perfiles probabilísticos o solamente perfiles.

Con la notación descrita, que son las mismas de ANADEP (Cordeiro, 1990), colocadas en la forma matricial se tiene: la matriz de probabilidades y los siguientes perfiles probabilísticos:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdot\cdot & \cdot\cdot & p_{1J} \\ p_{21} & p_{22} & \cdot\cdot & \cdot\cdot & p_{2J} \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ p_{I1} & p_{I2} & \cdot\cdot & \cdot\cdot & p_{IJ} \end{bmatrix}$$

$$P_{c.\pm} = \begin{bmatrix} p_{1\bullet} \\ p_{2\bullet} \\ \cdots \\ p_{I\bullet} \end{bmatrix} \qquad P_{r.\pm} = \begin{bmatrix} p_{\bullet 1} \\ p_{\bullet 2} \\ \cdots \\ p_{\bullet J} \end{bmatrix}.$$

Se define la matriz $\Delta$ , como

$\Delta = \frac{1}{\sqrt{2}}\left(\sqrt{P} - \sqrt{P_{c.\pm}}\sqrt{P'_{r.\pm}}\right)$ es una matriz de I filas con J columnas.

Usando los resultados anteriores, Cordeiro (1990) definió las Dependencias y Codependencias entre las columnas j y j$, respectivamente. Es decir:

$$d_{c.jj} = \frac{1}{2}\sum_{i=1}^{I}\left(\sqrt{p_{ij}} - \sqrt{p_{i\bullet}p_{\bullet j}}\right)^2$$

$$d_{c.jj\$} = \frac{1}{2}\sum_{i=1}^{I}\left(\sqrt{p_{ij}} - \sqrt{p_{i\bullet}p_{\bullet j}}\right)\left(\sqrt{p_{ij\$}} - \sqrt{p_{i\bullet}p_{\bullet j\$}}\right)$$

La siguiente matriz

$$D_C = \Delta'\Delta = \begin{bmatrix} d_{c,11} & d_{c,12} & \cdots & d_{c,1J} \\ d_{c,21} & d_{c,22} & \cdots & d_{c,2J} \\ \cdots & \cdots & \cdots & \cdots \\ d_{c,J1} & d_{c,J2} & \cdots & d_{c,JJ} \end{bmatrix}$$

contiene las dependencias y codependencias entre las columnas de la tabla de contingencia y se denomina matriz de dependencias y codependencias.

A partir de los valores de la matriz anterior se define la matriz

$$\Psi_C = \begin{bmatrix} 1 & \delta_{c,12} & \cdots & \delta_{c,1J} \\ \delta_{c,21} & 1 & \cdots & \delta_{c,2J} \\ \cdots & \cdots & \cdots & \cdots \\ \delta_{c,J1} & \delta_{c,J2} & \cdots & \delta_{c,JJ} \end{bmatrix}$$

que contiene los coeficientes de codependencia entre dos columnas cualesquiera de la matriz delta, donde:

$$\delta_{c,jj\$} = d_{c,jj\$} \Big/ \left(d_{c,jj} \quad d_{c,j\$j\$}\right)^{1/2}$$

Cabe resaltar que las matrices anteriores contienen las dependencias y codependencias en variables cualitativas, - que emulan varianzas y covarianzas de variables numéricas- y los coeficientes de codependencias- que emulan a los coeficientes de correlaciones en variables numéricas (Cordeiro, 1990).

Para facilitar la posterior presentación del Análisis de trayectoria con Variables Categóricas (Gómez, 1997), la matriz $\Delta$ es re-escrita de la siguiente manera:

$$\Delta = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,J-1} & x_{1,J} \\ x_{21} & x_{22} & \cdots & x_{2,J-1} & x_{2,J} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{I1} & x_{I2} & \cdots & x_{I,J-1} & x_{IJ} \end{bmatrix}$$

En éste punto, dependiendo de las relaciones existentes entre los niveles de las categorías de la tabla de contingencia se puede abordar el problema desde dos ópticas:

- En el primer caso, un nivel de una de las categorías de la tabla de contingencia es considerado como una variable respuesta, al que se denomina **nivel respuesta** y los otros niveles son tratados como variables explicativas y se denominan **niveles explicativos**. En éste caso el objetivo es estudiar la influencia marginal o combinada de los niveles explicativos en la distribución del nivel respuesta.
- En el segundo caso, el interés está dirigido al estudio de **asociación** entre los diversos niveles de una de las categorías de la tabla de contingencia, por ejemplo, para la verificación de existencia de algún grado de dependencia entre tales niveles. En éste caso, todos los niveles de la categoría son considerados como variables respuesta (niveles respuesta).

En el contexto descrito, el desarrollo del Análisis de Trayectoria para Tablas de Contingencia, se realiza utilizándose relaciones funcionales, o solamente llevando en cuenta relaciones de asociación entre los niveles de una de las categorias de la tabla de contingencia bidimensional.

En el presente artículo, se presenta un resumen de la propuesta metodológica del Análisis de Trayectoria en Tablas de Contingencia Bidimensional, para el primer caso. Se plantea la metodología para las columnas de la tabla de contingencia bidimensional de la misma manera que se haría para las filas.

**Método Propuesto de Análisis de Trayectoria**

Para la aplicación del análisis de trayectoria con variables numérica, inicialmente se considera un diagrama de trayectoria asociado a las variables del problema (Wright, 1918; Achcar, 1976; Dillon *et al.*, 1984; entre otros).

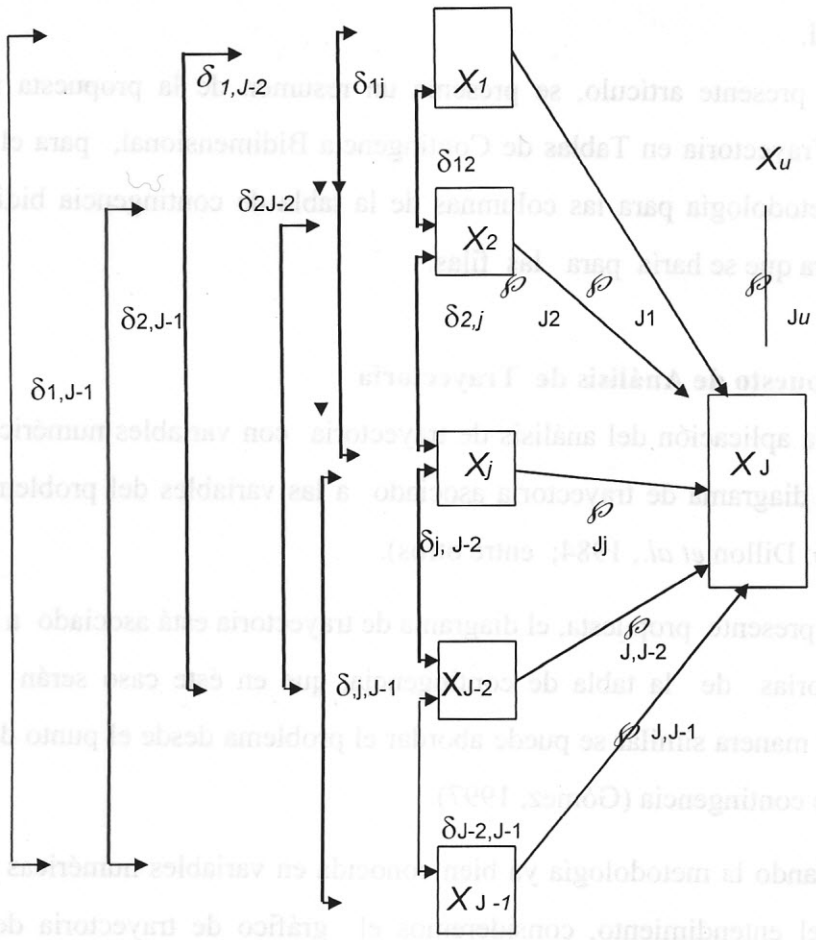
En la presente propuesta, el diagrama de trayectoria está asociado a los niveles de una de las categorías de la tabla de contingencia, que en éste caso serán los niveles de la columna. De manera similar se puede abordar el problema desde el punto de vista de las filas de la tabla de contingencia (Gómez, 1997).

Emulando la metodología ya bien conocida en variables numéricas y con la finalidad de facilitar el entendimiento, consideremos el gráfico de trayectoria de la Figura 1, las siguientes definiciones y las siguientes consideraciones:

- Bj son los niveles de la columna j=1,...,J, que resultan cuantificados mediante las columnas, Xj, de la matriz $\Delta$,
- $X_j$: j-ésima columna de la matriz $\Delta$,
- $_{ij}$ coeficientes de trayectoria , donde el subíndice "i" está asociado al nivel efecto o respuesta y el índice "j" está asociado al nivel causa o nivel explicativo. Mide el efecto directo del j-ésimo nivel de la variable columna sobre el i-ésimo nivel del mismo, llevando en consideración todos los niveles de las filas de la tabla de contingencia.

**FIGURA 1**

**Gráfico de trayectoria para el modelo propuesto**




- En correspondencia al diagrama de trayectoria asociamos a cada flecha unidireccional un coeficiente de trayectoria y a cada flecha bidireccional un coeficiente de codependencia,
- Al diagrama de trayectoria se asocia un conjunto de ecuaciones a las cuales denominamos "ecuaciones de regresión", donde cada nivel efecto será un nivel dependiente ( de los otros niveles) en el modelo de regresión propuesto (asociado al diagrama de trayectoria), donde:
- $\beta_{Jj\bullet l}$ coeficientes de regresión,
- $l = (1,2,\ldots, j-1, j = 1,\ldots, J-1)$.
- La relación entre un coeficientes de trayectoria, $_{Jj}$ y un coeficiente de regresión, $\beta_{Jj\bullet l}$, donde $l = (1,2,\ldots, j-1, j-1,\ldots, J-1)$ se obtiene escribiendo la ecuación de regresión en dos formas equivalentes.

- Así, para el gráfico de la Figura 1 se asocia el siguiente modelo de regresión múltiple en tablas de contingencia bidimensionales. Esta será considerada la primera forma :

$$X_{Ji}=\beta_{J1\bullet23\ldots J-1}X_{1i}\ \beta_{J2\bullet13\ldots J-1}X_{2i}\ \cdots\beta_{J,J-1\bullet13\ldots J-2}X_{J-1,i}\ \varepsilon_i$$

El modelo en la forma matricial resulta ser:

$$\vec{X}_J = \mathbf{X}\,\vec{\beta}\pm\vec{\varepsilon}$$

donde:

$\vec{X}_J = \begin{bmatrix} x_{1J} \\ x_{2J} \\ \cdots \\ x_{IJ} \end{bmatrix}$ es la J-ésima columna de la matriz $\Delta$ ,

$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,J-1} \\ x_{21} & x_{22} & \cdots & x_{2,J-1} \\ \cdots & \cdots & \cdots & \cdots \\ x_{I1} & x_{I2} & \cdots & x_{I,J-1} \end{bmatrix}$ es la matriz con las columnas restantes de la matriz $\Delta$ ,

$\vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_I \end{bmatrix}$ es el vector de errores, y

$\vec{\beta} = \begin{bmatrix} \beta_{J1\bullet234\ldots J-1} \\ \beta_{J2\bullet134\ldots J-1} \\ \cdots \\ \beta_{J,J-1\bullet12\ldots J-2} \end{bmatrix}$ es el vector de coeficientes de regresión y es el que deberá ser estimado

a través de la propuesta metodológica del presente trabajo.

De manera similar, a la Figura 1 se asocia el siguiente modelo de trayectoria o segunda forma:

$$Z_{Ji}=\wp_{J1}Z_{1i}\ \wp_{J2}Z_{2i}\cdots\wp_{J,J-1}Z_{J-1,i}\ \ \wp_{J,u}Z_{ui}$$

donde:

$$Z_{ji} = \frac{X_{ji}}{\sqrt{d_{c,jj}}} \qquad j= 1,2,..,J \; ; \qquad i = 1, 2,...,I$$

$$_{Ju}Z_{u} = \varepsilon_{J} .$$

El modelo en la forma matricial resulta ser:

$$\vec{Z}_J = \mathbf{Z}\ \vec{\wp} \pm \vec{\varepsilon}$$

donde:

$\vec{Z}_J = \begin{bmatrix} Z_{1J} \\ Z_{2J} \\ \cdots \\ Z_{IJ} \end{bmatrix}$ es la J-ésima columna de la matriz $\Delta$ estandarizada

$Z = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1,J-1} \\ Z_{21} & Z_{22} & \cdots & Z_{2,J-1} \\ \cdots & \cdots & \cdots & \cdots \\ Z_{I1} & Z_{I2} & \cdots & Z_{I,J-1} \end{bmatrix}$ es la matriz con las columnas restantes de la matriz $\Delta$ estandarizada,

$\vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_I \end{bmatrix}$ es el vector de errores, y

$\vec{\wp} = \begin{bmatrix} \wp_{J1} \\ \wp_{J2} \\ \cdots \\ \wp_{J,J-1} \end{bmatrix}$ es el vector de coeficientes de trayectoria y es el que también deberá ser estimado a través de la propuesta metodológica.

- Relacionando los dos modelos, los coeficientes de trayectoria (modelo en la segunda forma) se pueden escribir en función de los coeficientes de regresión(modelo en la primera forma). Así se llega a demostrar que:

$$_{Jj} = \beta_{Jj\bullet 12...(J-1)(J-1)} \left( d_{c,jj} / d_{c,JJ} \right)^{1/2}$$

y se propone estimar el vector de parámetros de los coeficientes de trayectoria mediante la siguiente fórmula, cuya demostración y detalles se encuentra en Gómez, 1997.

$$\hat{\vec{\ }} = (Z'Z)^{-1} Z' \vec{Z}_J$$

Para estudiar otras relaciones entre los coeficientes de trayectoria y coeficientes que emulen a los coeficientes de correlaciones parciales, correlaciones múltiples; pruebas de hipótesis para coeficientes de trayectoria; teorema fundamental del análisis de caminos; aplicaciones, etc.; en variables cualitativas, puede remitirse a Gómez, 1997.

## Conclusiones

- Fué posible hacer la transposición de la teoría de análisis de trayectoria en variables cuantitativas a tablas de contingencia bidimensionales.
- Con la transposición de la metodología de análisis de trayectoria para datos categóricos se demuestra la equivalencia entre los dos tipos de formulaciones para datos categóricos en tablas bidimensionales a través de relaciones funcionales; funciones a las que se han denominado de regresión en un caso y funciones de trayectoria en el otro caso.
- Se definen conceptos de coeficientes de codependencia parciales y múltiples y a partir de esas definiciones es posible emular definiciones para coeficientes de trayectoria en tablas de contingencia bidimensional.
- Se pueden descomponer los coeficientes de contingencia en tablas de contingencia, en sus efectos directos e indirectos.

## Referencia Bibliográfica


1. Achcar, J. Análise de trajetória. Sao Paulo, 1976. 68p. Disertaçao de Mestrado. Instituto de Matemática e Estatística, Universidade de Sao Paulo.
2. Agresti, A. Analysis of ordinal categorical data. New York: John Wiley, 1984. 287p.
3. Andersen, E. The statistical analysis of categorical data. New York: Springer Verlag, 1991. 531p.

4. Bishop, Y & Fienberg, S. & Holland, P. Discrete multivariate analysis: theory and practice. Cambridge: Mit Press, 1984. 557p.
5. cordeiro, j.a. Análise de Dependencia. Sao José de Rio Preto, 1990. 34p. Tese de Livre Docencia. IBILCE, Universidade do Estado de Sao Paulo.
6. Fienberg, S. E. The analysis of cross classified categorical data. Cambridge: The Mit Press, 1980. 198p.
7. Gómez, D. Análise de dependencia por trajetória. Sao Paulo, 1997. 121p. Disertaçao de Doutorado. ESALQ, Universidade de Sao Paulo.
8. Goodman, L.A. The analysis of cross classified data having ordered and or unordered categories: association models correlation models and analysis asymetrics models for contingency table with and without missing entries. The Annals of Statistics, v. 13, p. 10-69, 1985.
9. Li, C. Path analysis: a primer. Boxwood: Pacific Grover, 1975. 346p.
10. Wermuth, N. Linear recursive equations, covariance selection, and path analysis. Journal of the American Statistical Association, v. 75, n. 372, p. 963-972, 1980.
11. Wright, S. The method of path coefficients. The Annals Mathematics of Statistics, n. 5, p. 161-215, 1934.