

Clasificación mediante análisis de conglomerados: un método relevante para la psiquiatría.

Classification by cluster analysis: a relevant method for psychiatry.

Johann M. Vega-Dienstmaier^{1a}, J. Martín Arévalo-Flores^{1a}.

RESUMEN

El objetivo de este artículo es describir el método de análisis por conglomerados y sus aplicaciones especialmente en la investigación psiquiátrica. El análisis de conglomerados es un método que permite clasificar individuos u objetos en grupos homogéneos (conglomerados o clusters). Dentro de cada conglomerado los objetos tienen características similares entre ellos pero diferentes a las de los objetos de otros conglomerados. El análisis por conglomerados tiene aplicaciones en diversas áreas que van desde la biología hasta los negocios, particularmente es muy importante para la psiquiatría donde la clasificación de los trastornos mentales ha sido motivo de controversia.

PALABRAS CLAVE: Análisis de conglomerados, clasificación, psiquiatría.

ABSTRACT

The objective of this paper is to describe the cluster analysis method and its applications especially in psychiatric research. Cluster analysis is a method for classifying individuals or objects into homogeneous groups (clusters). Within each cluster, objects have similar features among them but different from those of objects of other clusters. Cluster analysis has applications in diverse fields ranging from biology to business, it is particularly important for psychiatry where the classification of mental disorders has been controversial.

KEYWORDS: Cluster analysis, classification, psychiatry.

INTRODUCCIÓN

El análisis de conglomerados es un método que tiene la finalidad de identificar grupos homogéneos de objetos, pacientes u observaciones (unidades de análisis), los cuales se denominarán conglomerados, grupos o clusters. Dentro de cada conglomerado los objetos tienen características similares entre ellos; pero diferentes a las de los objetos de otros conglomerados (1). El análisis por conglomerados permite, de este modo, clasificar a los individuos en base a sus características, lo cual puede tener

aplicaciones en diversas áreas que van desde la biología hasta los negocios, particularmente es muy importante en especialidades donde hay diversas propuestas sobre clasificaciones diagnósticas pero sin llegar claramente a un consenso, tal es el caso de la psiquiatría y otras disciplinas relacionadas con salud mental. Otro método relacionado con la clasificación es el análisis de función discriminante, sin embargo, mientras en este último se estudia la relación entre las características de los individuos y la clase a la que pertenecen (la cual ya se conoce), el análisis por conglomerados intenta descubrir las clases que están detrás de las características de los individuos (2-4).

¹ Universidad Peruana Cayetano Heredia, Sección de Psiquiatría y Salud Mental; Hospital Nacional Cayetano Heredia, Servicio de Neuro-Psiquiatría. Lima, Perú.

^a Psiquiatra.

El objetivo de este artículo es describir el método de análisis por conglomerados y sus aplicaciones especialmente en la investigación psiquiátrica.

MEDICIÓN DE LA SIMILITUD O DIFERENCIA

Uno de los primeros pasos para hacer el análisis de conglomerados es la medición del grado de similitud o diferencia entre los objetos que queremos clasificar (1).

Escojamos un aspecto (variable) para evaluar la similitud entre determinados objetos. Por ejemplo, la talla de un grupo de 6 personas que se muestra en la Tabla 1.

Tabla 1. Ejemplo para medición de distancias en base a las características físicas de 6 personas.

Persona	Talla (cm)	Peso (kg)
A	160	60
B	161	62
C	163	68
D	170	71
E	172	72
F	175	73

Aquí podemos observar que la talla de B difiere de la talla de A en 1 cm (161 – 160 cm), mientras que la diferencia de tallas entre F y A es de 15 cm (175-160cm); por lo tanto hay mayor similitud entre A y B, que entre A y F. Asimismo, hay mayor distancia entre C y F (12cm), que entre A y B (1cm).

La situación puede complicarse un poco cuando consideramos ahora las 2 variables (talla y peso). Graficando estas 2 variables se obtiene la Figura 1.

En base a la cercanía entre los puntos de este gráfico, se observa a simple vista que los primeros 2 sujetos (a la izquierda y abajo) forman un grupo, los 3 últimos (a la derecha y arriba) forman otro grupo, y hay un individuo aislado entre estos 2 grupos.

Hay varias formas de medir las distancias entre los puntos, la más usada para variables continuas es la distancia euclidiana (una línea recta entre 2 puntos); sin embargo, también tenemos la distancia “de cuadras de ciudad” o *city-block distance* (la suma de las diferencias absolutas de los valores de las variables) y la distancia de Chebychev (la máxima diferencia

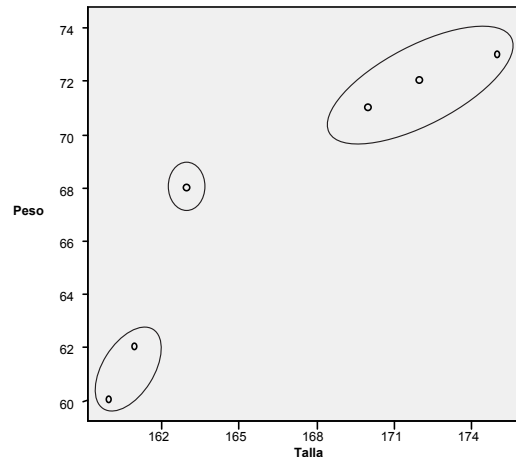


Figura 1. Gráfico de la talla y peso de 6 personas.

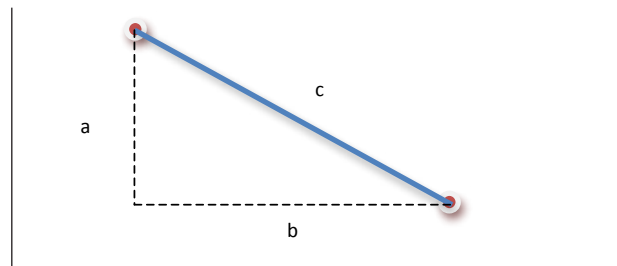


Figura 2. Formas de medir las distancias entre objetos. Euclidiana: c “City-block”: a+b Chebychev: b (dado que: b>a)

absoluta de los valores de las variables) (1). En la Figura 2 se puede apreciar estos conceptos.

Como las unidades de las variables pueden ser muy diferentes se puede estandarizar los valores de las variables antes del análisis, a través de diferentes métodos, tales como la transformación a los correspondientes valores z y la estandarización por rango (por ejemplo, a valores de 0 a 1 o -1 a +1).

A medida que aumentan las variables en base a las que se van a agrupar los objetos, los cálculos se complican cada vez más, pero mediante programas estadísticos se puede facilitar este trabajo.

Lo que hemos discutido hasta ahora se puede aplicar a variables numéricas, sin embargo, en caso de que las variables sean binarias o nominales, se emplearían medidas de similitud que expresan el grado en que las variables de 2 objetos concuerdan.

Por ejemplo, tenemos los objetos A, B, C... y J con 3 variables v1, v2 y v3 (Tabla 2).

Tabla 2. Ejemplo para medición de distancias considerando variables categóricas.

Objeto	v1	v2	v3
A	0	0	0
B	0	0	0
C	0	0	1
D	0	1	1
E	1	1	1
F	0	1	0
G	1	0	1
H	0	1	0
I	1	0	0
J	1	1	1

v1, v2, v3 = variable 1, variable 2, variable 3.

Podemos observar aquí que los valores de las 3 variables de A y B son los mismos, por lo que tienen 100% de concordancia; A y J no tienen ningún valor en común, por lo que su concordancia es 0; C y D concuerdan en 2 de las 3 variables (concordancia=0,667); y finalmente, la concordancia entre E y F es 0,333. Esta medida de similitud se denomina emparejamiento simple, pero existen muchas otras medidas de concordancia tales como la de Jaccard, de Russel & Rao, la Q de Yule, Kulczynski, Ochiai, Sokal & Sneath, Rogers & Tanimoto, Hamman y Dice, cuya forma de calcularse se muestra en la Tabla 3 (5).

Muchas bases de datos tienen a la vez variables nominales, ordinales y continuas, lo cual puede complicar el análisis. Una alternativa es dicotomizar las variables que no son binarias y usar las medidas de concordancia que acabamos de comentar. Otra opción es usar el método de análisis de conglomerados en 2 fases, que discutiremos más adelante.

Tabla 3. Medidas de concordancia o similitud para variables dicotómicas.

Medida de concordancia	Cálculo	Valores posibles
Concordancia simple	$(a+d)/(a+b+c+d)$	0→1
Russel & Rao	$a/(a+b+c+d)$	0→1
Jaccard	$a/(a+b+c)$	0→1
Dice o Czekanowski o Sorenson	$2a/(2a+b+c)$	0→1
Sokal & Sneath 1	$2(a+d) / (2(a+d)+b+c)$	0→1
Rogers & Tanimoto	$(a+d)/(a+d+2(b+c))$	0→1
Sokal & Sneath 2	$a/(a+2(b+c))$	0→1
Kulczynski 1	$a/(b+c)$	0→infinito
Sokal & Sneath 3	$(a+d)/(b+c)$	0→infinito
Ochiai 1	$a / \text{raíz}\{(a+b)(a+c)\}$	
Ochiai 2	$ad/\text{raíz}\{(a+b)(a+c)(b+d)(c+d)\}$	
Q de Yule	$(ad-bc)/(ad+bc)$	
Hamann	$\{(a+d)-(b+c)\}/(a+b+c+d)^2$	

Para evaluar similitud entre casos:		Objeto 2, características		a: número de características presentes en ambos objetos. b: número de características presentes en el objeto 1 pero ausentes en objeto 2 c: número de características presentes en el objeto 2 pero ausentes en objeto 1 d: número de características ausentes en ambos objetos.
		Presentes	Ausentes	
Objeto 1, características	Presentes	a	b	
	Ausentes	c	d	

Para evaluar similitud entre variables		Variable 2, objetos con el valor		a: número de objetos con el valor presente en ambas variables. b: número de objetos con el valor presente en la variable 1 pero ausente en la variable 2. c: número de objetos con el valor ausente en la variable 1 pero presente en la variable 2. d: número de objetos con el valor ausente en ambas variables.
		Presente	Ausente	
Variable 1, objetos con el valor	Presente	a	b	
	Ausente	c	d	

Así como podemos hacer análisis de conglomerados de objetos, podemos hacer lo mismo con las variables; en caso de variables binarias, las medidas de similitud serían las mismas que se describen en la Tabla 3; y en caso de variables numéricas, los coeficientes de correlación.

MÉTODOS DE ANÁLISIS DE CONGLOMERADOS

Existen 3 tipos de métodos para el análisis de conglomerados (1-4):

- Los métodos jerárquicos.
- Los métodos no jerárquicos
- El análisis en dos fases.

Los métodos jerárquicos

Los métodos jerárquicos forman una estructura en forma de árbol en el curso del análisis. La mayoría pertenecen a la categoría de la agrupación aglomerativa, es decir se parte fusionando los objetos más similares para formar pequeños conglomerados y progresivamente estos conglomerados se unen con otros objetos o conglomerados cada vez menos parecidos hasta terminar con los más diferentes juntándose en un solo grupo. Sin embargo, también hay métodos de agrupación divisiva, en los que se parte de todos los objetos en un solo conglomerado para luego irlos dividiendo. En los métodos jerárquicos, si un objeto es asignado a un determinado grupo no tiene posibilidad de ser reasignado a otro, a diferencia de los llamados métodos de partición.

Algoritmos de agrupamiento

Los algoritmos de agrupamiento dependen de la forma de definir la distancia entre un conglomerado y un objeto u otro conglomerado (Figura 3). Los más conocidos son:

Vinculación simple: la distancia entre 2 conglomerados corresponde a la menor distancia entre un miembro de un grupo y uno del otro grupo. Tiende a formar un conglomerado grande con los otros conglomerados formados por sólo 1 o pocos objetos. Puede ser útil para detectar objetos atípicos. Es considerado el algoritmo más versátil.

Vinculación completa: la distancia entre 2 conglomerados corresponde a la mayor distancia entre un miembro de un grupo y uno de otro grupo. Es afectada fuertemente por los objetos atípicos. Los conglomerados tienden a ser compactos y

estrechamente agrupados.

Vinculación por promedios: La distancia entre 2 conglomerados es el promedio de las distancias entre todos los pares formados por un miembro de un grupo y un miembro de otro grupo.

Vinculación por centroides: La distancia entre 2 conglomerados es la distancia entre los centroides (centros geométricos) de los grupos.

La vinculación por promedios y por centroides tienden a generar grupos de tamaños similares y con baja varianza dentro de cada conglomerado.

El proceso de vinculación sigue una serie de pasos en los que cada vez se une prioritariamente a los objetos y/o conglomerados que tienen menor distancia entre sí, hasta que todos los individuos estén en un mismo grupo. Este procedimiento se puede mostrar en un gráfico llamado dendrograma, en el cual se puede observar las distancias a nivel de las que ocurrió la fusión de objetos y grupos.

Vamos a mostrar un ejemplo usando la Tabla 2 de 10 objetos (A hasta J) y 3 variables binarias (v_1 , v_2 y v_3). Empleamos el programa estadístico SPSS para hacer un análisis de conglomerados jerárquicos midiendo la similitud entre objetos mediante concordancia simple, obteniéndose la matriz que se muestra en la Tabla 4. Asimismo, mostramos el dendrograma producto del análisis jerárquico utilizando el algoritmo de vinculación por centroides en la Figura 4.

El SPSS cuando grafica el dendrograma transforma las medidas de distancia o similitud a una escala de distancias de 0 a 25. Aquí se puede observar que E y J son muy cercanos, y efectivamente los 2 objetos tienen los mismos valores en cada una de sus variables, y lo mismo ocurre con A y B. Asimismo, se puede observar que B y E están muy alejados (ninguna de sus 3 variables concuerdan) y recién se fusionan a la derecha del dendrograma a la distancia máxima de 25.

Otra aproximación es el método de Ward, aquí se juntan los objetos cuya combinación aumenta la varianza dentro del grupo de manera mínima. Es el método preferido cuando no hay valores atípicos y se desea obtener grupos de similar tamaño.

Métodos no jerárquicos o de partición

El procedimiento de k-medias es uno de los más importantes métodos de partición. No vincula

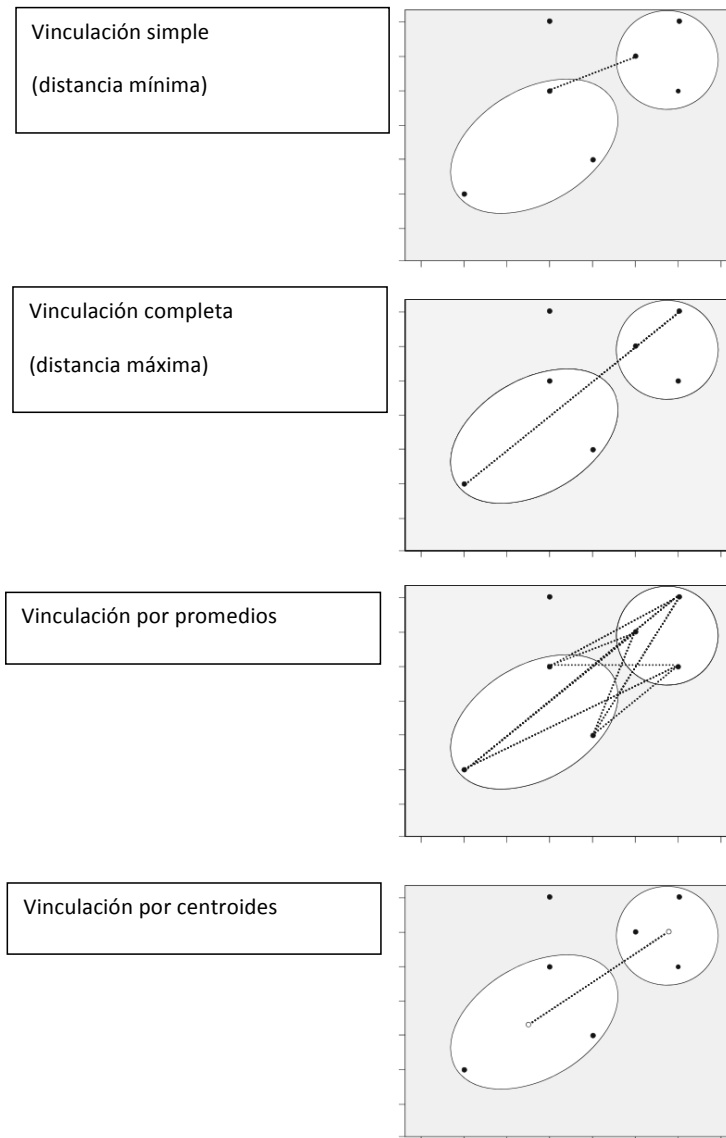


Figura 3. Algoritmos de agrupamiento.

Tabla 4. Matriz de similitudes de los objetos de la Tabla 2 por emparejamiento simple.

Caso	medida de emparejamiento simple									
	1:A	2:B	3:C	4:D	5:E	6:F	7:G	8:H	9:I	10:J
1:A	1,000	1,000	0,667	0,333	0,000	0,667	0,333	0,667	0,667	0,000
2:B	1,000	1,000	0,667	0,333	0,000	0,667	0,333	0,667	0,667	0,000
3:C	0,667	0,667	1,000	0,667	0,333	0,333	0,667	0,333	0,333	0,333
4:D	0,333	0,333	0,667	1,000	0,667	0,667	0,333	0,667	0,000	0,667
5:E	0,000	0,000	0,333	0,667	1,000	0,333	0,667	0,333	0,333	1,000
6:F	0,667	0,667	0,333	0,667	0,333	1,000	0,000	1,000	0,333	0,333
7:G	0,333	0,333	0,667	0,333	0,667	0,000	1,000	0,000	0,667	0,667
8:H	0,667	0,667	0,333	0,667	0,333	1,000	0,000	1,000	0,333	0,333
9:I	0,667	0,667	0,333	0,000	0,333	0,333	0,667	0,333	1,000	0,333
10:J	0,000	0,000	0,333	0,667	1,000	0,333	0,667	0,333	0,333	1,000

referencia el tamaño muestral sugerido para análisis de clases latentes (otro método para clasificar objetos diferente al del análisis por conglomerados), éste sería 2^k , en donde k es el número de variables que se van a utilizar para clasificar a los individuos (7).

VALIDACIÓN E INTERPRETACIÓN DE LOS RESULTADOS

Un aspecto deseable de la solución de conglomerados es que tenga estabilidad, es decir que se obtengan resultados similares ya sea utilizando diferentes métodos de análisis sobre los mismos datos, o aplicando el mismo método de análisis a cada una de las 2 mitades de la base de datos.

La interpretación de los resultados involucra describir y comparar las variables entre conglomerados. Por ejemplo, para el caso de variables continuas, se puede usar la prueba t o el análisis de varianza (ANOVA); si se encuentra diferencia significativa entre las medias de las variables de los conglomerados, éstos pueden distinguirse, lo cual respalda la solución hallada.

Finalmente podemos asignarle un nombre conveniente a cada conglomerado para identificarlo (1).

APLICACIONES DEL ANÁLISIS POR CONGLOMERADOS

El análisis por conglomerados tiene aplicaciones en diversos campos (4): En biología, tendría aplicaciones en la taxonomía de la diversidad de especies de seres vivos que existen en la naturaleza; en meteorología, tiene utilidad para descubrir los patrones relacionados con los cambios del clima en distintas regiones del mundo; en tecnologías de información, sirve para agrupar la infinidad de datos que se encuentran en Internet y así facilitar el acceso a ellos; en negocios, se usa para clasificar distintos segmentos de clientes y diseñar estrategias específicas de marketing para cada uno de ellos; y en medicina, puede servir para clasificar de una manera más adecuada síndromes y trastornos.

En cuanto a la psiquiatría, se han hecho duras críticas a los sistemas de clasificación de trastornos mentales. Se sostiene que los diagnósticos psiquiátricos presentan heterogeneidad, tienen una alta comorbilidad y no existen límites claros entre ellos, ni entre salud y enfermedad; que las categorías actuales ignoran aspectos importantes de los problemas de muchos

pacientes; que los criterios diagnósticos conducen a los estudiantes a ignorar fenomenología relevante; que las categorías no ayudan a probar hipótesis neurobiológicas de los trastornos mentales; que los criterios diagnósticos van cambiando sustancialmente a través de los años; y que las categorías son esencialmente producto de la opinión de expertos y no de la investigación científica (8,9). Considerando todos estos problemas, aplicar un método de clasificación como el análisis por conglomerados en psiquiatría tiene una gran relevancia.

A continuación describiremos algunos estudios recientes que se han hecho mediante análisis por conglomerados con pacientes psiquiátricos.

El análisis de conglomerados se ha utilizado para identificar tipos diferentes de pacientes con trastornos mentales severos. Considerando sexo, edad, funcionalidad, nivel educativo, el tipo de trastorno mental (trastornos afectivos, abuso de alcohol, esquizofrenia o trastornos de personalidad), la presencia de tendencias suicidas, el vivir de manera autónoma (o bajo supervisión) y la ayuda que se recibe por parte de los servicios de salud, es posible separar a los pacientes en 5 clases: 1) Mujeres mayores más funcionales con trastornos afectivos que reciben poca ayuda de los servicios de salud; 2) Hombres de edad intermedia con diversos trastornos mentales y abuso de alcohol que reciben ayuda insuficiente e inadecuada de los servicios de salud, 3) Mujeres de edad intermedia con necesidades serias, con trastornos afectivos, de personalidad y tendencias suicidas, que viven de forma independiente y que reciben ayuda amplia pero inadecuada; 4) Varones jóvenes con alto grado de instrucción con esquizofrenia que viven de manera independiente y que reciben ayuda adecuada; y 5) Varones mayores con bajo grado de instrucción, con esquizofrenia que viven con supervisión y ayuda percibida como adecuada. Esta clasificación sería útil en la planificación de los servicios de salud (10).

El clasificar a los pacientes en base a la sintomatología que tienen puede ser útil para predecir la respuesta al tratamiento farmacológico. Por ejemplo, haciendo un meta-análisis de los ensayos clínicos realizados con duloxetina para evaluar su eficacia para depresión mayor, se han identificado 5 clases según la sintomatología basal de los participantes: 1) Pacientes con falta de conciencia de enfermedad; 2) Participantes con alteraciones del sueño, disfunción sexual y síntomas somáticos; 3) Individuos con depresión mayor típica; 4) Participantes con síntomas gastrointestinales y disminución de peso; y 5)

pacientes con depresión mayor leve. En este estudio se encontró que el efecto del tratamiento fue mayor en el grupo con problemas del sueño, sexuales y somáticos; y en el de pacientes con síntomas gastrointestinales y disminución de peso (11).

Asimismo, la identificación de subgrupos de trastorno depresivo mayor mediante análisis por conglomerados, puede ser útil para predecir el pronóstico a largo plazo de los pacientes. Así tenemos que un estudio ha encontrado que el grupo de pacientes caracterizados por un episodio índice con disforia severa, tendencias suicidas, ansiedad e inicio temprano, es el de peor pronóstico (12).

En un estudio canadiense, mediante análisis de conglomerados, se encontró que los suicidios se pueden clasificar en 5 tipos tomando en cuenta determinadas características demográficas y clínicas, específicamente: el sexo, la edad, el estado civil (ser casado o no), si hubo evento estresante reciente, la enfermedad mental (depresión, trastorno bipolar, esquizofrenia, o si no había enfermedad mental conocida), la historia de abuso de sustancias, el antecedente de intento previo, el método empleado (si fue violento o no, si fue mediante salto desde altura) y el dejar nota suicida. De modo que las características del tipo 1 fueron: sexo femenino, tener depresión, el uso de métodos no violentos y haber tenido algún intento previo; las del tipo 2: ser casado, haber tenido un evento estresante reciente y el uso de métodos violentos; las del tipo 3: sexo masculino, haber tenido un evento estresante reciente, sufrir de enfermedad mental y tener historia de abuso de sustancias; las del tipo 4: no ser casado, ser joven, tener trastorno bipolar o esquizofrenia, e intento suicida por salto de altura; y las del tipo 5: no ser casado, no haber tenido intentos previos, no tener enfermedad mental previa conocida y dejar nota suicida (13).

Finalmente, usando análisis por conglomerados se ha encontrado 3 grupos de pacientes con juego patológico: El tipo I mostró menos psicopatología y rasgos de personalidad más funcionales; el II, problemas emocionales importantes, vergüenza, inmadurez, hostilidad y sentimientos negativos; y el III, las alteraciones psicopatológicas más severas y rasgos esquizotípicos (14).

CONCLUSIÓN

El análisis por conglomerados es un método con aplicaciones en diversos campos de la investigación

dentro y fuera de la medicina; específicamente en psiquiatría puede permitir el progreso de los sistemas de clasificación de los trastornos mentales.

Correspondencia:

Dr. Johann M. Vega-Dienstmaier

E-mail: johann.vega.d@upch.pe

Av. José Pardo 1142-701, Miraflores, Lima 18, Perú.

REFERENCIAS BIBLIOGRÁFICAS

1. Mooi E, Sarstedt A. Cluster analysis. En: A concise guide to market research. Berlin: Springer-Verlag; 2011.
2. Landau S, Everitt BS. Classification: Cluster analysis and discriminate function analysis; Tibetan Skulls. In: A handbook of statistical analyses using SPSS. Boca Raton: Chapman & Hall/CRC; 2004.
3. Green PE, Carmone FJ, Smith SM. Dimension reducing methods and cluster analysis. En: Multidimensional Scaling. Addison Wesley; 1989.
4. Tan P-N, Steinbach M, Kumar V. Cluster analysis: Basic Concepts and algorithms. In: Introduction to data mining. Addison-Wesley; 2006.
5. Choi S-S, Cha S-H, Tappert CC. A Survey of binary similarity and distance measures. Systemics, Cybernetics and Informatics. 2010; 8(1): 43-48.
6. Dolnicar S. A Review of unquestioned standards in using cluster analysis for data-driven market segmentation. CD Conference Proceedings of the Australian and New Zealand Marketing Academy Conference 2002(ANZMAC 2002). Melbourne: Deakin University; 2-4 December 2002.
7. Formann AK. Die latent-class-analyse: Einführung in die theorie und anwendung. Weinheim: Beltz; 1984.
8. Nesse RM, Stein DJ. Toward a genuinely medical model for psychiatric nosology. BMC Medicine. 2012, 10:5.
9. Faravelli C, Castellini G, Landi M, Brugnera A. Are psychiatric diagnoses an obstacle for research and practice? Reliability, validity and the problem of psychiatric diagnoses. The case of GAD. Clin Pract Epidemiol Ment Health. 2012; 8:12-15.
10. Fleury MJ, Grenier G, Bamvita JM, Tremblay J. Typology of persons with severe mental disorders. BMC Psychiatry. 2013;13:137. doi: 10.1186/1471-244X-13-137.
11. Schacht A, Gorwood P, Boyce P, Schaffer A, Picard H. Depression symptom clusters and their predictive value for treatment outcomes: Results from an individual patient data meta-analysis of duloxetine trials. J Psychiatr Res. 2014 Feb 12. pii: S0022-3956(14)00031-4. doi: 10.1016/j.jpsychires.2014.02.001. [Epub ahead of print]

12. van Loo HM, Cai T, Gruber MJ, Li J, de Jonge P, Petukhova M, Rose S, Sampson NA, et al. Major depressive disorder subtypes to predict long-term course. *Depress Anxiety*. 2014 Jan 14. doi: 10.1002/da.22233. [Epub ahead of print]
13. Sinyor M, Schaffer A, Streiner DL. Characterizing suicide in Toronto: an observational study and cluster analysis. *Can J Psychiatry*. 2014; 59(1):26-33.
14. Jiménez-Murcia S, Granero R, Stinchfield R, Fernández-Aranda F, Penelo E, Savvidou LG, Fröberg F, Aymami N, Gómez-Peña M, Moragas L, del Pino-

Gutiérrez A, Fagundo AB, Menchón JM. Typologies of young pathological gamblers based on sociodemographic and clinical characteristics. *Compr Psychiatry*. 2013; 54(8):1153-60.

Recibido: 07/02/2014

Aceptado: 17/03/2014