Psicothema
2000. Vol. 12, Supl. nº 2, pp. 393-395

ISSN 0214 - 9915 CODEN PSOTEG

Evaluación del sesgo en el análisis exploratorio de datos multivariados

Manuel Morales Ortiz, T. Jurado Muriel y M. L. López Domínguez
Universidad de Sevilla

El objetivo de la presente investigación fue evaluar si el sesgo cometido por los sujetos en una tarea de clasificación de estímulos varió en función del formato de presentación (soles o estrellas) y la forma de asignar las variables (aleatoria, ordenadas según su correlación en 360 grados o representadas en el espacio mediante un biplot: soles o estrellas factoriales). Se encontró que hubo una interacción significativa entre el formato de presentación y la forma de asignar las variables. En concreto, se obtuvo que los sujetos cometieron menos errores cuando clasificaron los soles factoriales que las estrellas factoriales. Asimismo, también clasificaron mejor los soles ordenados que las estrellas ordenadas. Tampoco hubo diferencias significativas entre soles y estrellas cuando no se incluyó ningún tipo de información acerca de las correlaciones (asignación aleatoria), ni entre los soles factoriales y los soles ordenados. Por último, también se obtuvo que los sujetos tardaron más tiempo en completar la tarea en la condición de asignación aleatoria que en la representación factorial.

Multivariate exploratory data analysis: bias assessment. The objective of this investigation was to evaluate if the bias of the subjects varied during a stimulus classification task in relation to the presentation format (suns and stars) and to the way the variables were assigned (randomly, ordered according to their correlation in 360º or using a biplot: factorial suns or stars). It was found that there was a significant interaction between the presentation format and the method of assigning variables. In particular, it was found that the subjects made fewer mistakes classifying the factorial suns than the factorial stars. Likewise, they also classified better the ordered suns than the ordered stars. There were no significant differences between the suns and stars when no information regarding the correlation was included (random order) nor between the ordered suns and factorial suns. Lastly, it was also found that the subjects took longer to complete a task with random assignment than one with the factorial representation.

Son varios los autores que han propuesto utilizar distintas técnicas gráficas para realizar análisis exploratorios de los datos multivariados (Wang, 1978;Toit, Stumpf & Steyn, 1986; Everitt, 1994), ya que su forma visual permite un mejor y más rápido procesamiento de la información. En la actualidad, el desarrollo de los recursos informáticos permite elegir entre distintas opciones (caras de Chernoff, gráficos de perfiles, soles, estrellas, etc.). Sin embargo, surge el problema de saber cuál de ellos es el que facilita la tarea del análisis de los datos. Algunos autores han mostrado sus preferencias por los gráficos de caras (Jacob, 1978; Borg & Staufenbiel, 1992), mientras que otros han mostrado su desacuerdo (Everitt, 1994).

Los trabajos que han estudiado este problema no presentan datos concluyentes. Así, el trabajo pionero de Mezzich y Worthington (1978) no encontró diferencias muy grandes entre los gráficos de caras y los de perfiles (62.9 % de aciertos frente al 60.5%). Sin embargo, en ese estudio no se incluyeron los gráficos de soles ni los gráficos de estrellas. Jacobs (1978) realizó una serie de experimentos en los que comparó distintos formatos de caras de Chernoff con matrices de dígitos y con estrellas. Encontró que los gráficos de estrellas eran mejores en una tarea de aprendizaje de pares asociados, mientras que los gráficos de caras redujeron significativamente el número de errores frente a los gráficos de polígonos y a las matrices de dígitos.

Existe alguna evidencia a favor de representar directamente la información relevante con objeto de facilitar el éxito en la realización de una determinada tarea (Cleveland, 1985; Spence & Lewandosky, 1990). Así, Mezzich & Worthington (1978), encontraron que los sujetos cometían menos errores en una tarea de clasificación cuando se les presentaba directamente los valores de los dos primeros factores en un espacio bidimensional. También encontraron que los sujetos realizaron mejor una tarea de clasificación con gráficos de fourier en los que se introducían las puntuaciones de los sujetos en los factores que con otros gráficos e incluso con gráficos de fourier en los que se introducían los datos de cada sujeto directamente. Asimismo, Lambertina, Freni-Titulaer y Louv (1984) compararon histogramas con gráficos de árboles y gráficos de castillos y encontraron que aquellos gráficos que mostraban directamente algún tipo de información correlacional (los gráficos de árbol), fueron los que produjeron el menor número de errores. Además, no encontraron diferencias entre los gráficos de histogramas con asignación aleatoria de las variables e histogramas que presentaban agrupadas las variables más correlacionadas. Por último, Borg y Staufenbiel (1992) encontraron que los sujetos realizaban mejor una tarea de clasificación con los gráficos de so-

Correspondencia: Manuel Morales Ortiz
Facultad de Psicología
Universidad de Sevilla
41005 Sevilla (Spain)
E-mail: morales@psicoexp.us.es

les factoriales (representaciones que proyectan las variables en un plano bidimensional definido por los dos factores principales), que con los gráficos de polígonos (estrellas) o con los soles.

En consecuencia, sabemos que la ejecución de los sujetos en tareas como las anteriormente mencionadas depende de presentar la información relevante directamente y del formato gráfico (p. ej. soles, gráficos de fourier). Sin embargo, los distintos trabajos que se han realizado bien no han tenido en cuenta estas dos variables o bien las han considerado de forma que resulta imposible separar el efecto de cada una de ellas. Por tanto, todavía no tenemos datos que hayan comparado los distintos formatos gráficos bajo dos condiciones: 1) sin presentar la información directamente, y 2) presentando la información directamente. A continuación, presentamos los resultados de un estudio en el que se tienen en cuenta estas dos variables.

Método

Sujetos

Se estudiaron de manera individual 34 estudiantes de Psicología de la Universidad de Sevilla. Ninguno de ellos tenía conocimiento previo sobre íconos. El experimento se explicó como un estudio de percepción.

Material

Se construyeron 44 gráficos de cada tipo, con los datos obtenidos por Mezzich y Worthington (1978) en un experimento donde 11 psiquiatras evaluaron el grado de aparición de determinados síntomas en 4 pacientes prototípicos con diferentes trastornos (depresivo, maníaco, esquizofrénico y paranoíde). Los soles simples y las estrellas fueron construidos usando un esquema en el que se dibuja cada variable (compuesta de valores entre 1 y 6) en su radio correspondiente. El orden de asignación de las variables fue aleatorio, situándose la primera variable en la zona superior al igual que las 12 en una esfera del reloj. Las siguientes variables fueron dispuestas con el mismo ángulo como si marcaran las distintas horas de un reloj. La misma estrategia se utilizó para construir los soles y las estrellas ordenadas, pero las variables fueron asignadas teniendo en cuenta la información obtenida sobre la estructura correlacional en el biplot. Así, en el lugar correspondiente a las 12 en la esfera del reloj se situó a la variable M y a continuación, se dispusieron las restantes variables siguiendo el orden de proximidad con dicha variable en el biplot. Para los soles factoriales se utilizó otro esquema de construcción basado en la proyección de las variables en un plano compuesto por los dos primeros componentes principales de sus intercorrelaciones. Por último, se construyó un tipo distinto de gráfico de estrella que consistió en unir cada una de las líneas de los gráficos de soles factoriales de modo que quedara un polígono cerrado al que nosotros llamamos «estrellas factoriales». Todos estos gráficos fueron presentados en folios de tamaño A4.

Procedimiento

Se utilizó un diseño factorial mixto 2 x 3 (gráfico x forma de presentar la información), donde la variable gráfico se estudió con una estrategia entresujetos. Los distintos gráficos se mostraron, de manera independiente, distribuidos aleatoriamente en una mesa de modo que todos ellos pudieran ser vistos simultáneamente por los sujetos. A cada sujeto se le pidió que examinara todos los gráficos y que los clasificara en cuatro grupos lo más parecidos posible. No se les puso un tiempo límite, pero sí se les indicó que prestasen más atención a la exactitud de la clasificación que a la rapidez. Se midió el número de aciertos y el tiempo que tardaron en realizar la tarea.

Resultados

Se realizó un ANOVA de medidas repetidas con cada una de las variables dependientes: 1) *Número de aciertos*. En la tabla 1 se presentan las medias y las desviaciones tipo para cada condición. Se encontró que la interacción entre gráfico y forma de presentar la información fue significativa ($F(2, 64) = 7.76$, $p < .01$). En concreto, se encontró que dentro de los soles hubo diferencias significativas entre la condición de soles aleatorios y soles ordenados ($t = -3.28$, $p < .01$), y entre soles aleatorios y soles factoriales ($t = -2.98$. $p < .01$). Además, no hubo diferencias significativas entre soles ordenados y soles factoriales ($t = 0.78$, $p = 0.45$). En cambio, dentro del grupo de gráficos de estrellas no hubo diferencias significativas entre ninguna de las condiciones.

2) *Tiempo de realización de la prueba*.- En la tabla 2 se presentan las medias y las desviaciones tipo para cada condición. Se encontró que el único efecto significativo fue el de la forma de presentar la información ($F(2, 64) = 4.63$, $p <.05$). Se encontró significativa la diferencia entre la asignación aleatoria y la factorial ($t = -2.88$, $p < .01$).

Tabla 1
Media y desviaciones de la variable número de aciertos

		Aleatoria	Ordenada	Factorial
SOLES	**MEDIA**	24	27.41	26.53
	SD	3.31	4.57	3.78
ESTRELLAS	**MEDIA**	25.18	23.65	22.76
	SD	5.77	3.71	3.27

Tabla 2
Media y desviaciones tipo de la variable tiempo de realización de la prueba

		Aleatoria	Ordenada	Factorial
SOLES	**MEDIA**	425.71	381.41	363.94
	SD	201.76	179.7	146.45
ESTRELLAS	**MEDIA**	327.35	311.47	261.94
	SD	146.3	133.03	96.8

Discusión

Los resultados de este estudio nos permiten extraer varias conclusiones:1) En todos los trabajos se pone de manifiesto que la información numérica (mediante dígitos) es la que presenta las mayores dificultades de procesamiento. 2) Los resultados confirman, en forma replicada, la hipótesis de que los soles factoriales son superiores a las estrellas. No obstante, la variedad de formatos que admite el gráfico de soles es bastante amplia, ya que en nuestro estudio no hemos encontrado diferencias significativas entre los soles ordenados y lo soles factoriales. 3) Hay diferencias significati-

vas en el procesamiento de los gráficos que no dependen del nivel de información que se presenta. Por tanto, sigue siendo necesario recurrir a la teoría psicológica para explicar las diferencias encontradas. Algunos autores han planteado la hipótesis de que en algunos gráficos se estén utilizando dimensiones más integrables (Garner, 1974).

Agradecimientos

Este trabajo fue financiado parcialmente por una beca PB93-1173 de la Dirección General de Investigación Científica y Técnica del Ministerio de Educación y Ciencia Español y por la Consejería de Educación y Ciencia de la Junta de Andalucía.

Referencias

Borg, I. & Staufenbiel ,T. (1992). Performance of Snow flakes, suns and factorial suns in the graphical representation of multivariate data. *Multivariate Behavioral Research, 27*(1), 43-55.

Cleveland, W.S. (1985). *The elements of graphing data*. Monterey: Wadsworth.

Everitt, B.S. (1994). Exploring multivariate data graphically: a brief review with examples. *Journal of Applied Statistics, 21*(3), 63-94.

Garner, W.R. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

Jacob, R.J.K.(1978). Facial representation of multivariate data. En P.C.C. Wang (Ed.). *Graphical representation of multivariate data*, (pp- 143-168). New York: Academic press.

Lambertina, W.J., Freni-Titulaer, W.J. & Louv, W.C. (1984). Comparisons of some graphical methods for exploratory multivariate data analysis. *The American Statician, 38*(3), 184- 188.

Mezzich, J.E. & Worthington, D.R.L. (1978). A comparison of graphical representations of multidimensional psychiatric diagnostic data. En P.C.C. Wang (Ed.). *Graphical representation of multivariate data*, (pp-123-141). New York: Academic press.

Spence, I. & Lewandowsky, S. (1990). Graphical perception. En J. Fox and J. S. Long (Eds.). *Modern methods of data analysis***,** (pp. 13-57). Newbury Park, CA: Sage.

Toit, S.H.; Stumpf, R.H. & Steyn, A.G.W. (1986). *Graphical exploratory data analysis.* Harrisonsburg: Donnelly and Sons.

Wang, P.C.C. (1978). *Graphical representation of multivariate data*. New York: Academic Press.