Una revisión a la clasificación de variables y su relación con la selección de su análisis estadístico

A review of classification of variables and its relationship with statistical analysis choice

Adriana Pérez M. (1)

CLASIFICACIÓN DE VARIABLES

Cuando se desea evaluar la solidez de la evidencia presentada en la literatura médica, se requiere que cuestionemos el tipo de análisis estadístico con el cual se validan los resultados presentados, es decir, ¿las pruebas estadísticas fueron apropiadas o se debieron realizar otro tipo de pruebas?

Para responder objetivamente a esta pregunta, se necesita caracterizar las variables de interés en el estudio. Entendiendo por variable cualquier característica que varía de un miembro a otro en una población determinada.

Para caracterizar las variables existen varias clasificaciones: según si hay interrupciones entre los valores observados de la variable, según su orientación descriptiva y según los niveles de medición (1,2).

Según si hay interrupciones entre los valores sucesivos de la variable se pueden clasificar en variables discretas y continuas (no existen interrupciones entre los valores).

Una variable discreta, por ejemplo, es el número de pacientes con asma, número de pacientes con enfermedad pulmonar obstructiva crónica (EPOC), etc. Una variable continua, por ejemplo, es la edad medida en años, meses, semanas y días, peso, talla, etc.

Según la orientación descriptiva, hace referencia a la situación donde la variable describe a otra o es descrita por otras variables.

Cuando la variable describe a otra variable se le considera una variable independiente y en el caso de que sea descrita por otras variables se le considera una variable dependiente. Esta clasificación depende de los objetivos del estudio y no de la naturaleza de la variable por sí misma.

Si una variable modifica la relación de otras dos variables y no es la variable principal del estudio, se identifica como una variable de confusión.

Según los niveles de medición, las variables se pueden clasificar en variables cualitativas, ya sean de tipo nominal u ordinal, o en cuantitativas, ya sean de intervalo o de razón (1,9).

Las variables nominales son aquellas que simplemente identifican categorías sin tener un orden, de acuerdo con algún criterio. Por ejemplo, el género de los pacientes, el color del esputo, etc.

Las variables ordinales nos permiten agrupar los valores de la misma en categorías, pero con la ventaja que siguen un orden específico. Por ejemplo, el grado de disnea de una persona, el tipo de neumonía, el grado de EPOC, el estado de carcinoma, etc.

La diferencia entre las variables de intervalo y de razón radica en el hecho de que el valor cero es arbitrario y absoluto respectivamente.

Un ejemplo de una variable continua de razón es la talla de la persona, la capacidad vital, el volumen residual, etc (no toman valores negativos ni exactos a cero).

(1) Epidemióloga egresada, Universidad Javeriana. Bogotá, Colombia.

Correspondencia*: Asociación Colombiana de Neumología y Cirugía de Tórax, Cra 16A Nº 80 - 74 Oficina 404 Bogotá, Colombia. Telefax: (571) 623 18 68 - 623 18 03*

Reimpreso de: *Rev. Colomb. Neumol. 2008; 8(4): 216-219.*

Un ejemplo de una variable continua de intervalo es la temperatura en grados Celsius.

Una variable puede ser clasificada en cualquiera de las tres formas, es decir, que no son mutuamente excluyentes. Por ejemplo, preclampsia para un estudio puede ser una variable dependiente, discreta y ordinal.

SELECCIÓN DEL ANÁLISIS ESTADÍSTICO APROPIADO

Cualquier investigador que se enfrenta al análisis de datos, se enfrenta a la selección del método estadístico. Para seleccionarlo, se deben tener en cuenta cuatro factores determinantes: el propósito de la investigación, las características matemáticas de la variable de interés, los supuestos estadísticos de las variables de interés y la forma en que los datos fueron recolectados.

Dentro de las características matemáticas de las variables el supuesto primordial para la elección de la técnica estadística esta dada por la distribución de la variable. Usualmente, se desea evaluar si la variable continua sigue la distribución normal (3) o guassiana o en la variable nominal, se desea evaluar si sigue una distribución binomial o se puede hacer aproximación a la norma debido a que el tamaño de la muestra es grande (mayor de 30 usualmente) (4).

A continuación aparecen enunciados los análisis estadísticos apropiados según las variables de interés y teniendo en cuenta cuántas muestras se relacionan en la etapa de análisis.

Las herramientas estadísticas aquí enunciadas hacen referencia a parámetros de uso común, como la media que identifica una medida de tendencia central y la varianza como una medida de dispersión, y no a otro tipo de pruebas como aquellas para bondad de ajuste, a intervalos de confianza, a distribuciones específicas, medidas de dependencia, etc.

UNA VARIABLE DE INTERÉS UNA SOLA MUESTRA

Teniendo en cuenta que la variables seleccionada se distribuyen normal, nos enfrentamos usualmente, a dos parámetros de interés, como son las varianza y/o la media.

Si estamos interesados en realizar pruebas inferenciales con respecto a la varianza, utilizaríamos una prueba chi-cuadrado para la varianza (5).

Si por el contrario, estamos interesados en realizar pruebas con respecto a la media, se puede pensar en una prueba t-student (1,4-6) cuando la varianza es desconocida. Cuando la varianza es conocida, la prueba apropiada sería una prueba normal (1,4-6).

Si la variable de interés se distribuye binomial, la prueba apropiada es la prueba binomial (5-7).

Si la variable de interés no se distribuye normal ni binomial, se necesitaría realizar pruebas no-paramétricas de acuerdo con su nivel de medición, así: nominal, prueba binomial (5-7); ordinal, prueba de cuarteles (7) y si es continua de intervalo, la prueba de Wilcoxon (5,7) es apropiada.

UNA VARIABLE DE INTERÉS DOS MUESTRAS

Teniendo en cuenta que la variable de interés se distribuye normal, nos enfrentamos con frecuencia a dos parámetros a comparar entre ambas muestras sus varianzas y/o sus medias.

Si deseamos comparar las varianzas de las muestras, se realiza una prueba F(5); si deseamos comparar las medias, se realizaría una prueba t (1,4-6) para varianzas desiguales o iguales si las muestras son independientes; en caso contrario se realizaría una prueba t pareada (5,6).

Si la variable de interés se distribuye binomial y las muestras son pareadas, el análisis es la prueba de McNemar (5,7). Si las muestras son independientes, la prueba adecuada es la binomial (5-7) o la prueba exacta de Fisher (6) (en caso de que los valores esperados sean menores de 6). Si los datos hacen referencia a incidencia, el análisis de incidencia (5,8) es el adecuado.

Si la variable de interés no se distribuye ni normal ni binomial, el análisis es no-paramétrico, dependiendo de la escala de medición y el tiempo de relación entre las muestras así: nominal, muestras pareadas, prueba de McNear(5,7) y para muestra independientes, la prueba chi-cuadrado (7); ordinal, muestras independientes, prueba de Mann-Withney (5,7) y para muestras pareadas, la prueba del signo (5,7); continua de intervalo con muestras independientes, prueba de aleatorización (7) y para muestras pareadas, prueba de Wilcoxon (5,7).

UNA VARIABLE DE INTERÉS MÁS DOS MUESTRAS

Si la variable de interés se distribuye normal y la inferencia a realizar es con respecto a las varianzas

se necesita realizar la prueba de Bartlett de homogeneidad de varianzas (5). Si la inferencia a realizar es con respecto a las medias, un análisis de varianza a ANOVA (5-7) o una prueba de Kruskal-Walls (5,7) es adecuada.

Si la variable de interés no se distribuye normal pero son datos categóricos, se pueden usar métodos para tablas de contingencia (5-7).

Si los datos no son categóricos, se pueden usar métodos apropiados para la distribución específica o métodos no-paramétricos, como la prueba chi-cuadrado (5-7), la prueba de la mediana (7) o la prueba de van der Waerden (7).

DOS VARIABLES DE INTERÉS AMBAS CONTINUAS

En el caso en que se deseen la predicción de una con la otra, el análisis apropiado es regresión lineal simple (1,5,6). Si se desea estudiar la relación entre ambas variables y ambas variables son normales, la técnica es el coeficiente de correlación de Pearson (1,5-7). Si alguna o las dos variables no son normales, con escala de medición por lo menos ordinal, se utilizan los coeficientes de correlación de Kendall (7) o Spearman (7).

DOS VARIABLES DE INTERÉS UNA CATEGÓRICA Y OTRA CONTINUA

El análisis recomendado en este caso es el ANOVA (5,6) si la variable continua se distribuye normal. Si la variable continua no se distribuye normal, la prueba de Kruskal-Walls (5-7) es la adecuada.

Si la relación de ambas variables debe ser controlada por otras variables estaremos enfrentándonos a posibles análisis de covarianza (1,9) ANOVA a dos vías o más (1,5,6), análisis de medidas repetidas (1.5.6), etc.

DOS VARIABLES DE INTERÉS AMBAS CATEGÓRICAS

En el caso de que ambas variables sean categóricas con escala de medición por lo menos ordinal, los métodos de correlación por rangos de Kendall (7) o Spearman (7) son adecuados.

Si ambas son categóricas nominales y estamos interesados en medidas de asociación, los métodos para tablas de contingencia (5-7) son adecuados. Si nuestro interés está en reproducibilidad, la técnica adecuada es la estadística Kappa (5,10).

MÁS DE DOS VARIABLES DE INTERÉS

En el caso en que las variables de interés sean continuas los métodos de regresión lineal múltiple (1,11) son apropiados, si se desea predecir una con respecto a las otras. Si se desea reducir el número de variables, el análisis de componentes principales (11) es el adecuado. Si se desean clasificar individuos dadas las variables, se utilizan los métodos jerárquicos (11) o el método de nubes dinámicas (11).

En el caso que la variable de respuesta es binomial y se desee establecer relaciones entre las variables donde el tiempo del evento es importante, el análisis adecuado es el análisis de sobrevida (12). Si el tiempo del evento no es importante, el análisis adecuado es el de regresión logística (13-15).

Si no se desea establecer relación entre las variables sino reducir el número de variables, los análisis de correspondencia binaria y múltiple (16) son adecuados para variables nominales y ordinales respectivamente.

Cuando se desean clasificar individuos y las variables son nominales, ordinales y continuas, el análisis discriminante (11) es el adecuado.

BIBLIOGRAFÍA

1. Kelinbaum DG, Kupper LL, Muler KE. Applied regression analysis and other multivariable methods. Belmont. Dusbury Press. 1988.
2. Rodríguez N, Pérez A. La bioestadística en la investigación clínica. Enviado para publicación. Medicina Interna. Editor: Chalem F, Escandon JE, Campos JY, Esguerra R De. Fundación Instituto de Reumatología e Inmunología. Tercera edición.
3. Shapiro SS, Wilk MB. Approximations for the null distribution of the W statistic. Tecnometrics1968; 10, 861-6.
4. Campbell MJ, Machin D. Medical Statistics. A commonsense approach. John Wiley & Sons. New York. Segunda Edición. 1993.
5. Rosner B. fundamentals of Biostatistics. Dusbury Press. Belmont. Tercera Edición. 1990.
6. Dawson-Sauders B, Trapp RG. Bioestadística Médica. Editorial El Manual Moderno S.A. México, 1993.
7. Conover WJ. Practical Nonparametric Statistics. John Wiley & Sons. 2nd Edition. New York. 1980.
8. Hennekens CH, Buring JE. Epidemiology in Medicine. Little Brown and Company. Boston, 1987.
9. Everitt BS, Statistical methods for medical investigators. Americas and John Wiley & Sons. 2nd edition. New york, 1994.
10. Kramer MS, Feinstein AR. Clinica Biostatistis. LIV. The Biostatistics of concordance. Clinical Pharmacology and Therapeutics 1981; 29: 111-23.
11. Jobson JD. Applied multivariate data analysis. Volumen II: Categorical and multivariate methods. Springer-Verlag. New York, 1992.

12. Kelinbaum DG. Survival analysis. A self-learning text. Springer-Verlag. New York, 1996.

13. Hosmer DW, Lemeshow S. Applied logistic regression. John Wiley & Sons. New York, 1989.

14. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. Am J Public Health 1991; 81: 1630-35.

15. Kleimbaum DG. Logistic regression: a self-learning text. Springer-Verlag. New York, 1994.

16. Pérez A. Análisis de correspondencia (binaria y múltiple) por medio del sistema de análisis estadístico (SAS). Tesis no publicada. Universidad Nacional de Colombia. Facultad de Ciencias. Departamento de Matemáticas y Estadística. Santa Fé de Bogotá, 1991.