ELECCIÓN DE UNA PRUEBA DE HIPÓTESIS

Jorge Dagnino S.[1]

- En la elección de una prueba de hipótesis se debe tomar en cuenta el diseño experimental, el tipo de distribución de la o las variables involucradas, la escala de medición y el número de variables o grupos estudiados.
- Se distingue entre pruebas paramétricas, basadas en la distribución normal y sus dos parámetros (media y varianza) y alternativas no paramétricas que no hacen presunciones sobre la distribución de la o las variables.
- No hay alternativas no paramétricas para el análisis de varianza de dos vías.

Hemos comentado el problema de las carencias en el conocimiento estadístico de los médicos y la necesidad de mejorar esa base. Por alguna razón, el tema de las dócimas no paramétricas (docimasia: conjunto de pruebas experimentales; dócima: prueba experimental) resulta un tema especialmente ajeno. Sin duda es un vacío en la formación de pregrado y, además, muchos de los textos clásicos le dan escasa o ninguna atención. Sin embargo, su uso es cada vez más frecuente en la literatura y los errores más frecuentes se relacionan con la descripción de datos ordinales como si fuesen de intervalo y la aplicación de pruebas de intervalo a datos ordinales. Resulta entonces particularmente importante discutir algunos conceptos relativos a las pruebas no paramétricas. La mayoría de los textos y cursos de estadística ponen énfasis en las pruebas mismas más que en los factores que llevan a elegir una de ellas. La disponibilidad de computadores y programas estadísticos ha hecho perder importancia al cómo hacer una prueba, trasladándola a la selección de ellas en el caso del investigador y a juzgar si la selección fue apropiada en el caso del lector. El objetivo de este artículo es discutir someramente los factores que influyen sobre la decisión de qué prueba usar para describir una muestra o para probar una hipótesis.

La inferencia estadística (derivar como consecuencia, conclusión o probabilidad), como hemos dicho, usa fundamentalmente dos herramientas: estimación de los parámetros de la población y pruebas de hipótesis. En el desarrollo de la estadística, las primeras técnicas que se desarrollaron fueron aquellas que hacían una serie de presunciones sobre la naturaleza de la o las poblaciones de las cuales provenían las muestras. Como los valores poblacionales se denominan "parámetros", estas técnicas se denominan "paramétricas". Por definición, producen resultados cuya confiabilidad está condicionada a que se cumplan las presunciones hechas al momento de diseñar las pruebas. En general, aquellas pruebas con las mayores presunciones son también las de mayor potencia (la capacidad de arrojar significativa una diferencia determinada con un n menor, o con un n igual una diferencia menor, que cuando se usa una prueba de menor potencia). El ejemplo más conocido, más usado y más mal utilizado, es la prueba t de Student, pero se puede mencionar también el análisis de varianza, la correlación de Pearson o la regresión lineal.

Las presunciones más importantes que deben ser satisfechas para usar las pruebas paramétricas son las siguientes:

1) Las observaciones deben ser independientes. Ello implica que la selección de cualquier individuo en la población de estudio no puede influir sobre las probabilidades de inclusión de cualquier otro, y el puntaje asignado a un individuo u observación no puede influir sobre el puntaje asignado a otro individuo o medición.
2) Las observaciones deben provenir de poblaciones con distribución Normal.
3) Estas poblaciones deben tener una varianza semejante.
4) Las variables implicadas deben haber sido me-

[1] Profesor Titular, División de Anestesiología, Pontificia Universidad Católica de Chile.

didas en por lo menos una escala de intervalos de manera que sea posible usar las operaciones aritméticas.

5) En el caso del análisis de varianza, además debe cumplirse que las medias de las poblaciones Normales y homoscedásticas deben ser combinaciones lineales de los efectos debidos a columnas o líneas. Dicho de otro modo, los efectos deben ser aditivos.

Después de las pruebas paramétricas, se desarrollaron técnicas que no hacen presunciones tan extensas o estrictas sobre los parámetros. Estas técnicas son denominadas "no paramétricas", de rango o de distribución libre. Sus resultados son condicionales a un menor número de calificadores que las pruebas paramétricas.

Para seleccionar el método estadístico más apropiado para conseguir estos fines, se deben tener en cuenta los siguientes factores: las características de la o las variables estudiadas (tipo y número), el tipo de datos y las escalas de medición.

El **tipo** de variables y su distribución: una variable es una característica que se mide en un estudio y puede ser independiente o dependiente. Variable independiente es aquella que está bajo el control del operador, la que es cambiada por éste, mientras que la variable dependiente es aquella no controlada, la que depende o es determinada por las variaciones de la o las variables independientes. Cuando dos o más variables independientes provienen de los mismos sujetos se habla de muestras o datos relacionados o pareados; cuando se obtienen de sujetos diferentes se habla de muestras no pareadas.

El **número** de variables independientes es otra de las determinantes de la elección del método estadístico apropiado. Por ejemplo, para estimar el riesgo anual de hipertermia maligna en un hospital, sin tener en cuenta las característica de los individuos, se habla de una variable y se usan métodos conocidos como análisis univariantes, que se aplican a una serie de observaciones que contienen una variable dependiente (hipertermia maligna) y ninguna independiente. Para examinar el riesgo de hipertermia maligna luego del uso de succinilcolina se debe usar un análisis bivariante, que se usa con grupos de observaciones con una variable dependiente y otra independiente (succinilcolina). Para evaluar el riesgo de hipertermia maligna en individuos con distrofia muscular, uso de halogenados y de succinilcolina, se usan los métodos de análisis multivariante, que se utilizan para grupos de observaciones que constituyen una variable dependiente (que sigue siendo la hipertermia maligna) y más de una independiente. Los métodos multivariantes se aplican con frecuencia para ajustar o despejar la influencia de variables de confusión. La decisión sobre cuál es la variable dependiente y cuál la independiente depende de la pregunta que se desea responder; dependiente es aquella que necesitamos valorar y la independiente es la condición que puede influir sobre la dependiente. Por ejemplo, si nos interesa saber si la respuesta a la succinilcolina es diferente en pacientes con hipertermia maligna, esta pasa a ser la variable independiente y aquella la dependiente.

El **tipo de escala de medición** indica la cantidad relativa de información que contiene cada una de ellas. Las mediciones de un nivel de información concreto pueden transformarse o reescalarse a un nivel inferior, pero no es posible reescalar las variables a un nivel superior al que se midieron realmente. Al reescalar a un nivel inferior se pierde información, hecho que tiende a aumentar el error tipo II, si todo lo demás se mantiene igual. En otras palabras, reescalar a un nivel inferior reduce la potencia estadística. Se dice que una prueba tiene una elevada potencia cuando tiene una escasa probabilidad de rechazar la hipótesis nula cuando esta es verdadera, pero una gran posibilidad de rechazarla cuando esta es falsa.

Normalmente las condiciones para hacer un análisis paramétrico (con la posible excepción de la homoscedasticidad) no se comprueban. Cuando tenemos razones fundadas para presumir que estas condiciones no se cumplen es imposible decir cual es realmente la potencia de la prueba y por lo tanto tampoco podemos hacer un pronunciamiento sobre las probabilidades en torno a la hipótesis planteada. Así, una probabilidad estimada de 0,04 puede en realidad ser de 0,035 o de 0,1.

Existe evidencia empírica para mostrar que "leves" desviaciones en las presunciones pudieran no tener efectos radicales sobre los resultados de probabilidad, pero no hay consenso en cuanto a qué constituyen desviaciones "leves". En el ámbito anestesiológico, una de las controversias se ha centrado en el tratamiento estadístico de la escala visual análoga. Si bien es efectivo que esta podría ser el mejor ejemplo de lo que es una "leve" desviación de las presunciones, vale la pena destacar que la pérdida de potencia al usar una prueba no paramétrica es escasa y que normalmente lo único que se necesita es aumentar el tamaño de la muestra para que las potencias sean equivalentes. Desde el punto de vista práctico, vale la pena decir que deben mirarse con cautela, siempre y no sólo por esta razón, resultados con una probabilidad demasiado cerca del 0,05 (o 0,01 según sea el caso). Cuando se tienen dudas o cuando el número de casos es peque-

ño, la conducta probablemente debiera ser conservadora, teniendo presente que se aumenta el error tipo II.

Al usar test paramétricos cuando el nivel de medición no lo justifica, se está añadiendo información al asignarle a los resultados cualidades que no tienen. A la inversa, al usar una prueba no paramétrica cuando está indicada una paramétrica se está perdiendo información.

Las ventajas de las pruebas no paramétricas se pueden resumir en los siguientes puntos:

1) Las probabilidades calculadas por la mayoría de las pruebas no paramétricas son probabilidades exactas (excepto en el caso de muestras muy grandes en las que existen excelentes aproximaciones) sin importar la distribución de la población desde la cual se tomó la muestra.
2) Con muestras tan pequeñas como n = 6, no hay alternativa a una prueba no paramétrica a menos que la naturaleza de la distribución de la población se conozca con exactitud.
3) Resistencia a los valores extremos o disparados ("outliers"). Estos no afectan a la mediana, pero pueden cambiar en forma importante la media.
4) Existen pruebas no paramétricas adecuadas para manejar muestras provenientes de observaciones de diferentes poblaciones. Ninguna de las pruebas paramétricas permite hacer esto sin necesidad de hacer presunciones poco realistas.
5) Son más fáciles de usar y aprender. El uso ubicuo de la computación eliminó completamente esta ventaja.

Las desventajas son fundamentalmente dos:

1) Pérdida de potencia cuando se cumplen todas las presunciones necesarias para aplicar una prueba paramétrica. El grado de pérdida de información se expresa en la relación potencia/eficiencia. Al decir que una prueba no paramétrica tiene una potencia/eficiencia de 90%, quiere decir que, de cumplirse todas las presunciones necesarias para aplicar una prueba paramétrica, esta sería tan efectiva como la no paramétrica con una muestra que es 10% menor que la usada en el análisis no paramétrico. De hecho, la prueba de Wilcoxon o la U de Mann-Whitney tienen una potencia/eficiencia que supera el 95% en relación a la prueba t de Student pareada y no pareada respectivamente.
2) No existen todavía pruebas no paramétricas para realizar el mismo tipo de análisis que se puede llevar a cabo con el análisis de varianza que permita dilucidar la interacción de factores.

En la Tabla 1 se resumen las pruebas estadísticas a elegir de acuerdo con el nivel de medición y el tipo de muestras. La tabla no es completa pues en ella se cita las pruebas más frecuentemente usadas y no todas las alternativas.

Tabla 1. Esquema de elección de acuerdo con el nivel de medida (Modificada de Siegel)

	Estadísticas descriptivas	**Una muestra**	**Dos muestras**		**n muestras**		**Correlación**
			Pareadas	Independientes	Pareadas	Independientes	
Nominal	Modo Frecuencias acumulativas	Binomial	McNemar	Fisher exacto	Cochran Q	χ^2 para n muestras	Coeficiente contingencia
Ordinal	Mediana Distribución frecuencias	Kolmogorov-Smirnov 1 muestra	Signo Wilcoxon	χ^2 Mann-Whitney U Kolmogorov-Smirnov 2 muestras	Friedman	Kruskal-Wallis	Spearman Kendall
Intervalo	Media DS Rangos		t pareado	t no pareado ANOVA 1 vía	ANOVA muestras repetidas	ANOVA 2 vías	Pearson

REFERENCIAS

1. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall,1991.
2. Bland M. An Introduction to Medical Statistics. 3rd Ed, Oxford: OUP, 2006.
3. Dawson-Saunders B, Trapp RG. Bioestadística médica. México D.F: Manual Moderno, 1993.
4. Dexter M. Wilcoxon-Mann-Whitney Test Used for Data That Are Nor Normally Distributed. Anesth Analg 2013; 117: 537-538.
5. Divine G, et al. A Review of Analysis and Sample Size Calculation Considerations for Wilcoxon Tests. Anesth Analg 2013: 117: 699-710.
6. Forrest M, Andersen B. Ordinal scale and statistics in medical research. Br Med J 1986; 292: 537-538.
7. Glantz SA. Primer of Biostatistics. 3a edición, New York: McGraw-Hill, 1992.
8. Ortega-Benito JM. Skewed distributions and parametric tests. Br Med J 1991; 303: 58.
9. Philip BK. Parametric statistics for evaluation of the visual analog scale. Anesth Analg 1990; 71: 708-713.
10. Portney LG, Watkins MP. Foundations of Clinical Research. Applications to practice. 2nd ed., Upper Saddle River: Prentice-Hall, 2000.
11. Riegelman RK, Hirsch RP. Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica. 2a ed., Washington, D.C.: OPS, 1992.
12. Siegel S. Non parametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.
13. Windish DM, Diener-West M. A clinician-educators' roadmap to choosing and interpreting statistical tests. J Gen Intern Med 2006; 21: 656-660.

Correspondencia a:
Dr. Jorge Dagnino S.
jdagnino@med.puc.cl