

Utilización de métodos robustos en la estadística inferencial

E. Ramalle-Gómara^a y J.M. Andrés de Llano^b

La estadística es la ciencia que permite tomar decisiones en situaciones de incertidumbre¹. Estas decisiones se basan con frecuencia en inferir a partir de muestras, tanto para estimar valores en las poblaciones como para realizar pruebas de contraste de hipótesis. Básicamente, el propósito de las pruebas de hipótesis es obtener conclusiones sobre los parámetros de la población (media, proporción u otros) basándonos en los resultados obtenidos en muestras aleatorias². La estadística ha desarrollado dos grupos de pruebas para tomar decisiones de este tipo: pruebas paramétricas y pruebas no paramétricas.

Las pruebas paramétricas tienen en cuenta los parámetros en las poblaciones. Para su utilización es necesario que se de una serie de requisitos o supuestos que, en caso de no cumplirse, impiden su utilización^{3,4}: la escala de medida debe ser de intervalo o razón; las observaciones deben ser independientes unas de otras; las variancias poblacionales de las variables en estudio deben ser similares (homocedasticidad); la relación entre las variables debe ser de tipo lineal, y la distribución de las variables en la población debe seguir una ley normal.

Las pruebas no paramétricas no necesitan estas condiciones previas de aplicación. Pueden utilizarse para analizar variables nominales y ordinales. La distribución poblacional puede ser cualquiera y no es ne-

cesario conocer ni suponer nada acerca de las variancias poblacionales. La mayoría no requiere el supuesto de linealidad.

Cuando se dan las condiciones de aplicación, las pruebas paramétricas tienen más potencia que las no paramétricas, pero, cuando esto no es así, el riesgo alfa puede ser mayor que el especificado de antemano⁵, es decir, se aumenta la probabilidad de afirmar una diferencia entre grupos que no existe, que podríamos traducir como que se incrementa la probabilidad de cometer un falso positivo⁶.

Una estrategia posible sería utilizar siempre pruebas no paramétricas ya que, si se dan las condiciones de aplicación, la pérdida de potencia no es muy grande y, si no se dan, son los métodos que deben emplearse. Es conocido que los métodos no paramétricos tienen una alta potencia cuando se dan las condiciones de aplicación de las paramétricas y tienen muy pocas probabilidades de conducir a una conclusión distinta de la obtenida por los métodos tradicionales paramétricos⁷.

Otra alternativa es la utilización de los llamados métodos robustos. Estos métodos son menos potentes que los paramétricos, pero se muestran superiores a los no paramétricos clásicos. Entendemos por potencia o poder de un test la probabilidad de elegir la hipótesis alternativa cuando es cierta, es decir, la capacidad para encontrar diferencias significativas cuando es cierto que existen^{8,9}.

La principal de sus ventajas es que no se afectan por la existencia de datos anómalos, como, por ejemplo, algunos datos muy extremos (*outliers*), y que no requieren los supuestos de aplicación de las pruebas paramétricas.

Cuando una distribución cualquiera presenta datos anómalos, existe el hábito generalizado de eliminarlos, proceso eufemísticamente llamado de limpieza o depuración, antes de realizar inferencias con ella. Esta actitud asume que los datos

LECTURA RÁPIDA

▼ La estadística es la ciencia que permite tomar decisiones sobre los resultados en situaciones de incertidumbre.

▼ El propósito de las pruebas de hipótesis es obtener conclusiones sobre los parámetros de la población (media, proporción u otros) basándonos en los resultados obtenidos en muestras aleatorias.

▼ Las pruebas paramétricas tienen en cuenta los parámetros en las poblaciones. Para su utilización es necesario que se de una serie de requisitos o supuestos.

▼ Las pruebas no paramétricas no necesitan estas condiciones previas de aplicación.

▼ Cuando se dan las condiciones de aplicación, las pruebas paramétricas tienen más potencia que las no paramétricas, pero, cuando esto no es así, el riesgo alfa puede ser mayor que el especificado de antemano.

^aServicio de Epidemiología. Gobierno de La Rioja. Logroño. España.

^bServicio de Pediatría. Hospital Río Carrión. Palencia. España.

Correspondencia:
Enrique Ramalle-Gómara.
Servicio de Epidemiología.
C/ Villamediana, 17.
26071 Logroño, La Rioja. España.
Correo electrónico: ramalle@larioja.org

Manuscrito recibido el 22 de julio de 2002.
Manuscrito aceptado para su publicación el 22 de julio de 2002.

LECTURA RÁPIDA

Una estrategia posible sería utilizar siempre pruebas no paramétricas ya que, si se dan las condiciones de aplicación, la pérdida de potencia no es muy grande y, si no se dan, son los métodos que deben emplearse.

Otra alternativa es la utilización de los llamados métodos robustos. Estos métodos son menos potentes que los paramétricos, pero se muestran superiores a los no paramétricos clásicos.

La principal de sus ventajas es que no se afectan por la existencia de datos anómalos.

Se asume que los datos extremos son erróneos, lo cual no es admisible. Un enfoque más adecuado es comprobar la veracidad de los datos. Si el dato extremo está equivocado debe corregirse.

Se puede sustituir por el valor promedio, por la interpolación de un dato con respecto a los adyacentes u otros métodos similares a los empleados cuando existen datos ausentes.

**TABLA
1**

Estimador	Estrategia	Resultado con los datos (1, 2, 3, 4, 5, 100) (media aritmética: 19,2)
Media α -winsorizada muestral	Se sustituye un determinado porcentaje, α , (20% generalmente) de valores extremos a cada lado de la muestra por el valor más próximo no sustituido	3,5
Media α -recortada muestral	Se eliminan las k observaciones extremas de cada lado, en lugar de winsorizarlas, calculando la media aritmética de las observaciones restantes	3,5
Mediana muestral	Divide la distribución en dos partes con el mismo número de elementos	3,5
Estimador de Huber	Se encuentra dentro de los denominados M-estimadores, que generalizan al estimador de máxima verosimilitud con buenas propiedades de robustez y eficiencia. En este caso se descartan las observaciones que sean mayores (o menores) a una constante	3,57

extremos son erróneos, lo cual no es admisible. Otras opciones que se utilizan son las de su sustitución por el valor promedio, por la interpolación de un dato con respecto a los adyacentes u otros métodos similares a los empleados cuando existen datos ausentes (*missing*) en una base de datos^{10,11}. Sin embargo, un enfoque más adecuado es comprobar la veracidad de los datos. Si el dato extremo está equivocado debe corregirse. Si es correcto, eliminarlo o sustituirlo puede modificar las inferencias que se realicen a partir de esa información, debido a que introduce un sesgo que es difícil de cuantificar¹² y a que disminuye el tamaño muestral. Este último aspecto es crucial cuando la muestra estudiada es pequeña. Los métodos robustos pueden ser de utilidad para la realización de inferencias sin tener que «depurar» los datos extremos, ya que están diseñados para realizar inferencias sobre el modelo, reduciendo la posible influencia que pudiera tener la presencia de datos anómalos¹³.

Métodos robustos para estimar medidas de centralización

Un problema conocido en el cálculo de medidas de posición o centralización es aquel en el que una distribución de datos no sigue una ley normal; en esta circunstancia la media no es un buen estimador del promedio de los datos. La media es particularmente sensible cuando la serie de datos es pequeña y existe algún valor extremo. Por ejemplo, en la serie 2, 3, 4, 5, 100, la media es 22,8, que no refleja bien el va-

lor promedio de la serie. La alternativa que suele proponerse a esta situación es utilizar la mediana y los percentiles para describir la distribución^{14,15}, ya que estas medidas no se ven afectadas, generalmente, por la existencia de valores extremos. La mediana de la distribución es 4. Sin embargo, en algunas circunstancias la mediana tampoco estima bien el promedio de la distribución. Esto ocurre cuando en una muestra la mediana y los valores superiores a ella están muy cercanos entre sí y, a su vez, muy alejados de los valores que se sitúan por debajo de la mediana, o viceversa. Por ejemplo, en la muestra 1, 2, 98, 99, 100, la mediana es 98, que no es un índice de resumen apropiado. El porcentaje máximo de valores extremos que soporta un estimador antes de no ser válido se llama punto de ruptura (*breakdown point*)¹⁶.

En estas circunstancias, tamaño muestral pequeño y presencia de valores extremos, pueden utilizarse métodos robustos para el cálculo de estadísticos de centralización o localización (tabla 1).

Como parámetros de localización se recomienda en primer lugar, el de Huber y, en segundo lugar, la media α -recortada muestral (con $\alpha = 0,2$).

Métodos robustos para estimar medidas de dispersión

Las medidas de dispersión clásicas (variancia y desviación típica) se ven afectadas por las mismas limitaciones que las medidas de posición. La desviación típica sólo es un buen estimador del promedio de la desvia-

TABLA	Métodos robustos para el contraste de hipótesis
3	

TABLA 4 Ejemplo para la comparación de dos medias

K-S-L: prueba de Kolmogorov-Smirnov, con corrección de Lilliefors para la comprobación del supuesto de normalidad.

Ejemplo para la comparación de dos medias
Disponemos de 74 pacientes ingresados en una unidad de cuidados intensivos (UCI) y queremos comparar las estancias generadas en dicha unidad entre un grupo de pacientes fallecidos y otro grupo de supervivientes (tabla 4).

Con los parámetros antes referidos, se pueden construir intervalos de confianza robustos y realizar contrastes de hipótesis¹³. Las principales pruebas de contraste de hipótesis basadas en métodos robustos se presentan en la tabla 3.

Las medidas de dispersión clásicas (variancia y desviación típica) se ven afectadas por las mismas limitaciones que las medidas de posición.

LECTURA RÁPIDA



Como alternativa, se recomienda la utilización de la desviación absoluta mediana estandarizada.



Métodos robustos para el contraste de hipótesis

Con los parámetros antes referidos, se pueden construir intervalos de confianza robustos y realizar contrastes de hipótesis.



El test de Yuen es una alternativa a las pruebas de la t de Student y U de Mann-Whitney para tamaños muestrales pequeños y distribuciones no normales.



Ejemplo para la comparación de más de dos medias

Para comparar más de dos medias, cuando no puede utilizarse el ANOVA clásico, deberíamos recurrir a los clásicos no paramétricos de Kruskal-Wallis o, mejor, a métodos robustos como la generalización robusta del test Welch, que utiliza medias α -recortadas muestrales.

TABLA 6

Grupo	Días de estancia	Media	Mediana	K-S-L
Jóvenes (n = 28)	8, 30, 55, 4, 3, 1, 1, 2, 2, 1, 2, 2, 1, 3, 1, 1, 2, 3, 2, 1, 2, 1, 3, 4, 1, 2, 3, 4	5,18	2,00	p < 0,0001
Maduros (n = 23)	1, 1, 1, 25, 13, 25, 4, 5, 4, 2, 1, 2, 33, 3, 2, 3, 2, 3, 1, 66, 11, 1, 3	9,22	3,00	p < 0,0001
Ancianos (n = 23)	1, 15, 25, 1, 31, 2, 11, 88, 3, 21, 60, 3, 1, 5, 13, 2, 1, 2, 1, 3, 27, 3, 1, 27	15,48	3,00	p = 0,0012

K-S-L: prueba de Kolmogorov-Smirnov, con corrección de Lilliefors para la comprobación del supuesto de normalidad.

TABLA 7

Resultados de las distintas pruebas para la comparación de más de dos medias

Prueba de contraste de hipótesis	Nivel de significación
Test de ANOVA	0,08
Test de Kruskal-Wallis	0,09
Método robusto media (α -recortada)	0,02

La variable estancia, como casi todas las que se refieren al tiempo, es una variable conocida por su distribución alejada de lo normal. Si el tamaño de los grupos que se estudian es pequeño —menor de 30—, no sería adecuada la utilización de métodos paramétricos (t de Student) y deberíamos recurrir a los clásicos no paramétricos (U de Mann-Whitney). Sin embargo, en este caso concreto ninguna de estas dos opciones tiene potencia suficiente para detectar diferencias significativas. La potencia es la capacidad para encontrar diferencias significativas cuando es cierto que existen⁸. Sin embargo, el test robusto de Yuen, que utiliza medias α -recortadas muestrales, es capaz de detectar diferencias significativas entre ambos grupos (tabla 5).

Por tanto, el test de Yuen es una alternativa a las pruebas de la t de Student y U de Mann-Whitney para tamaños muestrales pequeños y distribuciones no normales.

Ejemplo para la comparación de más de dos medias

Como ejemplo de la utilidad de los métodos robustos para el análisis de la variancia continuamos utilizando a otros 73 pacientes ingresados en una UCI y queremos comparar las estancias generadas en dicha unidad entre tres grupos de pacientes clasificados en función de la edad (jóvenes, maduros y ancianos). Las características que los describen se resumen en la tabla 6. Por otro lado, el test de Levene para valorar la homogeneidad de las variancias (homocedasticidad) muestra que existen diferencias entre ellas (p = 0,027). Por lo tanto, existen varios incumplimientos (ausencia de normalidad y homocedasticidad, y existencia de valores anómalos *outliers*) que impiden la utilización de métodos paramétricos (ANOVA clásico) en este ejemplo y deberíamos recurrir a los clásicos no paramétricos de Kruskal-Wallis o, mejor, a métodos robustos como la generalización robusta del test Welch, que utiliza medias α -recortadas muestrales y es capaz de detectar diferencias significativas entre los grupos (tabla 7).

Ejemplo para el análisis de correlación y regresión lineal

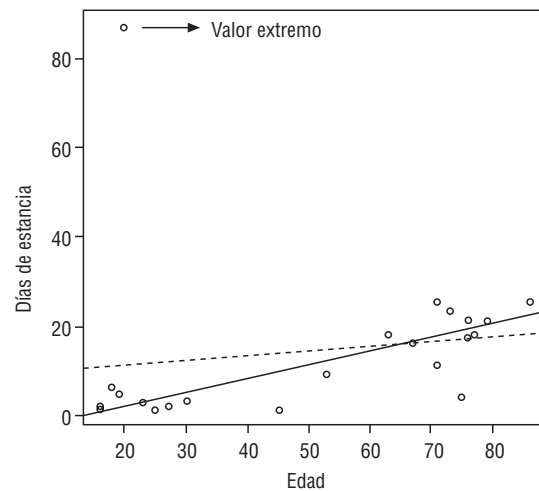
Por último, para comparar los distintos métodos en los análisis de correlación y re-

TABLA 8

Ejemplo para regresión lineal

Variable	Días de estancia	Media	Mediana	K-S-L
X (edad) (n = 22)	63, 79, 53, 20, 23, 18, 19, 16, 45, 30, 16, 67, 71, 73, 71, 76, 77, 75, 27, 86, 76, 25	50,27	58,00	p = 0,04
Y (estancia) (n = 22)	18, 21, 9, 87, 3, 6, 5, 2, 1, 3, 1, 16, 25, 23, 11, 21, 18, 4, 2, 25, 17, 1	14,50	10,00	p = 0,002

K-S-L: prueba de Kolmogorov-Smirnov, con corrección de Lilliefors para la comprobación del supuesto de normalidad.



Línea continua: ajuste mediante el estimador robusto de regresión medio bponderado
Línea de trazos discontinuos: ajuste mediante mínimos cuadrados

TABLA 9 Resultados para regresión lineal

Prueba de contraste de hipótesis	Coefficiente de correlación	Nivel de significación
Pearson	0,15	0,50
Spearman	0,61	0,002
Porcentaje ajustado poblacional	0,70	0,0003

gresión utilizaremos también a 22 pacientes ingresados en una UCI. En este caso deseamos valorar la relación entre la edad de los mismos y sus estancias generadas en dicha unidad (tabla 8).

De nuevo observamos la ausencia de normalidad en las dos variables que invalida la utilización de métodos paramétricos (correlación de Pearson y regresión lineal), por lo que deberíamos recurrir a métodos no paramétricos (Spearman) o, mejor, a métodos robustos como el coeficiente de porcentaje ajustado poblacional y el estimador robusto de regresión medio bponderado (tabla 9).

En la figura 1 podemos apreciar la diferencia entre la recta de regresión obtenida por el método de mínimos cuadrados y la obtenida por el estimador robusto de regresión medio bponderado, y cómo afecta la existencia de un valor extremo.

Como puede comprobarse en la figura 1, el ajuste es mejor con el estimador robusto que con la estimación por mínimos cua-

drados. El valor extremo apenas afecta a la estimación robusta.

La realización de cualquiera de estas estimaciones requiere un programa estadístico apropiado. Entre ellos puede utilizarse el programa R, que añade al hecho de facilitar estos cálculos el de ser de libre distribución, por lo que no está sometido a derechos de *copyright* y se obtiene gratuitamente en Internet¹⁷.

Bibliografía

1. Bailar JC III. Some ures of statistacal thinking. En: Bailar JC III, Mostoller F, editors. Medical ures of statistics. Boston: N Engl J Med Books, 1992; p. 25-57.
2. Guyatt GH, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Hypothesis testing. CMAJ 1995;152:27-32.
3. Esbrí R. El porqué de las asunciones de las pruebas paramétricas (distribución normal e igualdad de varianzas). Jano 1996;50:1120.
4. Manzano V. Inferencia estadística. Aplicaciones y usos con SPS/PC+. Madrid: RA-MA, 1995.
5. Siegel S, Castellan NJ. Estadística no paramétrica. 4.^a ed. México: Trillas, 1995.
6. Ramalle-Gómara E, Bermejo R. El significado de lo significativo. Algunas consideraciones sobre los tests de significación y el uso del valor «p». Aten Primaria 1994;14:863-5.
7. Colton T. Estadística en medicina. Barcelona: Masson-Little Brown, 1995.
8. Argimon JM, Jiménez-Villa J. Métodos de investigación clínica y epidemiológica. 2.^a ed. Barcelona: Harcourt, 2000.

LECTURA RÁPIDA

▼
Deberíamos recurrir a métodos no paramétricos (Spearman) o, mejor, a métodos robustos como el coeficiente de porcentaje ajustado poblacional y el estimador robusto de regresión medio bponderado.

▼
La realización de cualquiera de estas estimaciones requiere un programa estadístico apropiado. Entre ellos puede utilizarse el programa R.



9. Porta M, Moreno V, Sanz F, Carné X. Una cuestión de poder. *Med Clin (Barc)* 1989; 92:223-8.
10. Brick JM, Kalton G. Handling missing data in survey research. *Stat Methods Med Res* 1996;5:215-38.
11. Little RJ. Methods for handling missing values in clinical trials. *J Rheumatol* 1999;26: 1654-6.
12. Katz MH. *Multivariable analysis: a practical guide for clinicians*. Cambridge: Cambridge University Press, 1999.
13. García A. *Métodos avanzados de estadística aplicada*. Madrid: UNED, 2001.
14. Altman DG. *Practical statistics for medical research*. London: Chapman and Hall, 1991.
15. García M, Martín M. La media $\pm$ la desviación típica no siempre funciona. *Jano* 1996; 50:1508.
16. García A. *Métodos avanzados en estadística aplicada II. Métodos robustos y de remuestreo*. Madrid; 2001.
17. Disponible en: <http://www.R-project.org/>. 2002