Artículo original

Pruebas estadísticas en el cuidado de heridas

Palabras clave pruebas estadísticas inferenciales, valor p , intervalos de confianza, proporción de probabilidades

Como referencia Stephenson J. Statistical testing in wound care. WCET® Journal 2022;42(4):38-40

DOI https://doi.org/10.33235/wcet.42.4.38-40

Presentado el 6 de noviembre de 2022, Aceptado el 1 de diciembre de 2022

Los estudios sobre el cuidado de heridas suelen arrojar datos cuantitativos; y esos datos requieren algún tipo de análisis. Casi todos los estudios incluyen un resumen descriptivo de los participantes, como ya comenté en mi anterior artículo[1]. Sin embargo, muchos estudios también requieren algún tipo de pruebas estadísticas inferenciales, normalmente si se pretende generalizar los resultados de la muestra a una población.

Diferentes diseños de estudio requieren diferentes métodos de prueba, aunque el objetivo básico en la mayoría de los casos es el mismo. Se trata de evaluar la *importancia* del *efecto* de interés, es decir, determinar si el efecto que observamos en los datos de la muestra es un reflejo de la situación de la población de origen o si es probable que no sea más que una variación aleatoria de los datos. Un efecto puede ser una diferencia observada entre grupos de estudio (como la cantidad de biopelícula presente en un grupo de control y un grupo en el que los pacientes son tratados con un apósito antimicrobiano), la diferencia entre una medida tomada al inicio y después de la intervención de un único grupo de pacientes (como los niveles de dolor durante y después de la cirugía), una relación observada entre dos variables (como el grado de uso de tecnología sanitaria móvil y los conocimientos sobre el cuidado de heridas) o muchas otras cantidades.

El establecimiento de la significación estadística requiere la comprobación de una hipótesis. Normalmente se comprueba una *hipótesis nula* (de ausencia de efecto); por ejemplo, que la diferencia de medias en la población es cero. Miller et al[2] realizaron un ensayo controlado aleatorio (RCT) piloto simple ciego para examinar la concordancia con la terapia de estimulación eléctrica (EST) y su aceptabilidad en pacientes con úlceras venosas de la pierna (VLU) que no habían tolerado una compresión de moderada a alta. El tratamiento se contrastó con la hipótesis nula de ausencia de efecto; es decir, que la concordancia con el tiempo total de tratamiento recomendado bajo tratamiento de control o bajo EST era la misma. Lenselink y Andriessen[3] midieron (entre otras cantidades) el porcentaje de tejido de granulación y tejido amarillo en un estudio de cohortes sobre la eficacia de un apósito de biocelulosa con polihexanida en 28 pacientes, poniendo a prueba varias hipótesis relacionadas con las diferencias en los resultados de los pacientes entre el inicio y las 24 semanas. En un estudio de pacientes diabéticos y no diabéticos, Gunes et al[4] analizó varios biomarcadores, principalmente la galectina-3 (que promueve la angiogénesis y la formación de nuevos vasos) y probó varias hipótesis relativas a las relaciones entre la galectina-3 y varios otros biomarcadores.

La significación se cuantifica mediante el conocido (aunque no muy comprendido) valor *p*, que es una *probabilidad condicional*: la probabilidad de que se hubieran obtenido los resultados observados, o algo más extremo, dado que la hipótesis nula es cierta. En el lenguaje cotidiano, se denomina de forma general " la posibilidad de una oportunidad ": es decir, la probabilidad de que nuestros datos hayan caído de la forma en que lo han hecho sólo como resultado de una variación natural y no porque exista realmente un efecto. Convencionalmente, se considera que un valor *p* inferior a 0,05 (5%) indica un resultado de significación estadística (al nivel de significación del 5%) y el consiguiente rechazo de la hipótesis nula de ausencia de efecto. El estudio de Miller et al arrojó un valor *p* de 0,671 para su resultado primario, lo que indica evidencias insuficientes de una diferencia entre los grupos de tratamiento. El estudio de Lenselink y Andriessen arrojó un valor *p* inferior a 0,04 para una prueba que comparaba el porcentaje medio de tejido de granulación desde el inicio hasta 24 semanas después del tratamiento, lo que indica la existencia de un efecto del tratamiento. El estudio de Gunes et al arrojó un valor *p* inferior a 0,001 en una prueba de correlación entre los biomarcadores galectina-3 y proteína C reactiva, lo que indica una fuerte evidencia de una relación significativa entre estos biomarcadores.

El RCT realizado por Miller et al y la prueba de cambios con respecto al valor inicial realizada por Lenselink y Andriessen son ejemplos de diseños de estudios agrupados: un *estudio no emparejado*, en el que se comparan dos grupos de estudio no relacionados, en el caso de Miller et al y un diseño de *estudio emparejado* , en el que las medidas se toman en dos ocasiones del mismo grupo de pacientes, en el caso de Lenselink y Andriessen (la palabra "emparejado" se refiere al hecho de que cada paciente aporta un par de lecturas, no a que se trate de un par de grupos). Otra variante de este estudio surge cuando se toman medidas de pacientes emparejados en variables demográficas o relacionadas con la salud. El estudio de Gunes et al es un ejemplo de diseño correlacional que utiliza datos no agrupados. Estos tres diseños son probablemente las opciones más comunes de diseño de estudio en el cuidado de heridas, aunque pueden encontrarse varios otros tipos de estudio, que también pretenden inferir resultados de datos de muestra a una población, incluyendo por ejemplo pruebas de acuerdo, estudios de cribado/diagnóstico, análisis de supervivencia y metaanálisis.

John Stephenson
PHD FRSS (GradStat) CMath(MIMA)
Profesor titular de estadística biomédica
Universidad de Huddersfield, Reino Unido
Correo electrónico J.Stephenson@hud.ac.uk

Los resultados numéricos de los estudios no apareados pueden compararse mediante la prueba t de muestras independientes o, si intervienen más de dos grupos, una prueba como el análisis de la varianza (ANOVA). Los resultados categóricos suelen analizarse mediante la prueba de asociación chi-cuadrada. Para el diseño de estudio emparejado, la técnica de análisis estándar es la prueba t de muestras emparejadas o el ANOVA de medidas repetidas. Los estudios correlacionales suelen cuantificarse mediante el coeficiente de correlación de Pearson, y pueden ampliarse a un entorno de regresión lineal. Todos estos procedimientos son *paramétricos* y se basan en determinados supuestos de distribución de los datos; si no se cumplen estos supuestos, las técnicas no paramétricas correspondientes, como la prueba de suma de rangos, la prueba de rangos con signo, la evaluación del coeficiente de correlación de rangos de Spearman o los procedimientos de arranque pueden funcionar a menudo bastante bien. Todos los procedimientos pueden aplicarse fácilmente utilizando software estadístico estándar y todos darán lugar a una evaluación de la significación estadística medida por el valor *p* y dada por el software.

Mientras el valor *p* facilita una inferencia de significación estadística o en caso contrario, no nos da una medida de la precisión de nuestros resultados. Esta es otra cara de las evidencias inferenciales. Podemos encontrar en nuestra muestra un efecto de un tamaño determinado, pero no sería realista esperar que en la población en general exista un efecto precisamente de esa magnitud. Aunque nunca sabremos cuál es realmente el tamaño del efecto poblacional, es posible derivar un *intervalo de confianza* (CI), también conocido como *intervalo de incertidumbre*, para el mismo. A menudo se interpreta vagamente como el intervalo de valores dentro del cual podemos afirmar con un cierto grado de confianza (convencionalmente el 95%) que se encuentra un valor de la población. Más formalmente, si repitiéramos el procedimiento muchas veces, el intervalo de valores determinado cada vez contendría el verdadero parámetro poblacional en el 95% de las ocasiones.

Los CI no cuantifican la fuerza de la evidencia contra la hipótesis nula, como hace el valor *p*, pero sin embargo dan una medida de la precisión de una estimación (por ejemplo, la diferencia entre, o la proporción de, los valores medios en los grupos de tratamiento). En la actualidad, los CI (y la estimación del efecto en torno al cual se ajustan) se presentan cada vez más junto con los valores *p* en las pruebas estadísticas y a muchos les parecen más informativos y fáciles de entender. Aunque no forman parte de una prueba de hipótesis, la mayoría de los programas estadísticos informan automáticamente de un CI como subproducto de la salida de la prueba de hipótesis.

Existe una correspondencia exacta entre los CI y el valor p correspondiente: un CI del 95% que excluye el valor clave 0 (para una diferencia entre grupos de estudio) o 1 (para una proporción entre grupos de estudio) corresponde a un efecto significativo al nivel de significación del 5% (es decir, un valor *p* inferior a 0,05). En consecuencia, un CI del 95% que incluye un valor clave corresponde a un valor *p* que no es estadísticamente significativo al nivel de significación del 5% (es decir, es 0,05 o superior). Atkinson et al[5] investigaron el efecto de diversos factores sobre el riesgo de infección del sitio quirúrgico durante la cirugía de la columna vertebral y, en una presentación típica de resultados tabulados (abajo), informaron de las estadísticas de un modelo que incluía tanto valores *p* como CI. Obsérvese que el factor niveles medulares, que es significativo según el valor *p* (0,019), tiene una proporción de probabilidades con un CI asociado de 1,04 a 1,54, que excluye el valor clave (para una proporción) de 1; mientras que el factor región medular, que no es significativo según el valor p (0,103), tiene una proporción de probabilidades con un CI asociado de 0,71 a 44,3, que incluye el valor clave. Esta tabla también es un buen ejemplo de cómo un efecto de magnitud relativamente pequeña (cada nivel espinal adicional se asocia a un aumento del 26% en las probabilidades de infección) puede ser significativo; mientras que un efecto de gran magnitud (la cirugía realizada en la región torácica, en lugar de la no torácica, se asocia a un aumento de aproximadamente cinco veces en las probabilidades de infección) puede no ser significativo.

La selección de una prueba estadística adecuada para un determinado diseño de estudio no siempre es sencilla y requiere una cuidadosa consideración de los parámetros del estudio. Ninguna prueba es adecuada para todos los tipos de estudios. En el caso de los estudios agrupados, como los diseños no apareados y apareados comentados anteriormente, puede que tengamos que considerar, por ejemplo, el número de grupos, el tamaño de los grupos, la distribución de los datos, la independencia de las unidades y la presencia o ausencia de factores de confusión a la hora de seleccionar una prueba. En el caso de los estudios no agrupados, como los estudios correlacionales, es posible que debamos considerar si nuestros datos se presentan en forma de datos brutos u ordenaciones de rango; y en el caso de los estudios no agrupados que evalúan múltiples factores, es posible que debamos considerar si potencialmente estamos sobreajustando nuestros datos (modelando ruido en lugar de señal) al intentar analizar demasiados factores para una muestra de un tamaño determinado.

Se necesitan medidas especiales para diseños complejos: por ejemplo, cuando los datos están agrupados (como pacientes dentro de salas de hospital, o sitios anatómicos dentro de pacientes); cuando se realizan múltiples evidencias (como puede ocurrir en estudios con múltiples medidas de resultado, cuando la variable clave del tratamiento tiene múltiples niveles, o cuando se realizan análisis separados en subgrupos de individuos y/o en múltiples puntos temporales); cuando los resultados no se conocen con exactitud (como cuando un resultado es el tiempo transcurrido hasta un acontecimiento adverso en pacientes que se controlan con poca frecuencia); cuando se realiza una serie de observaciones en los mismos pacientes; o cuando es necesario determinar una estimación sintetizada del efecto a

Tabla 1: extracto de la tabla de datos comunicada por Atkinson et al[5]

Factor/covariable	valor p	Proporción de probabilidades	CI del 95% para proporción de probabilidades
Número de niveles de la columna vertebral	0,019	1,26	(1,04, 1,54)
Región vertebral primaria no torácica (referencia)			
Torácica	0,103	5,59	(0,71, 44,3)

partir de múltiples estudios. Ciertas características de los datos, como la presencia de muchos valores omitidos o no válidos, o de valores atípicos, también pueden requerir procesos estadísticos adicionales. En todas estas situaciones se recomienda solicitar el asesoramiento de un estadístico biomédico.

REFERENCIAS

1. Stephenson, J. (2022). Descriptive presentation of wound care data. World Council of Enterostomal Therapists Journal. 42, 3, p. 30-33
2. Miller C, McGuiness W, Wilson S, Cooper K, Swanson T, Rooney D, Piller N, Woodward M (2017). Concordance and acceptability of electric stimulation therapy: a randomised controlled trial. Journal of Wound Care Vol. 26, No. 8. https://doi-org.libaccess.hud.ac.uk/10.12968/jowc.2017.26.8.508
3. Lenselink E, Andriessen A (2011). A cohort study on the efficacy of a polyhexanide-containing biocellulose dressing in the treatment of biofilms in wounds. Journal of Wound Care 20 (534)
4. Gunes EA, Eren MA, Koyuncu I, Taskin A, Sabuncu T (2018). Investigation of galectin-3 levels in diabetic foot ulcers. Journal of Wound Care (27); 12
5. Atkinson R, Stephenson J, Jones A, Ousey K. An assessment of key risk factors for surgical site infection in patients undergoing surgery for spinal metastases J Wound Care 2016; 25(S9); S30-S34