

Tanto va el cántaro a la fuente...

...que termina por romperse. ¿Qué se rompe, el cántaro o la fuente?. El refrán se refiere, claro está, al cántaro. El dicho hace referencia a los tiempos en que no había agua en las casas y había que ir hasta la fuente a por ella, de forma que, más tarde o más temprano, el cántaro se rompía, ya fuese por desgaste por un uso excesivo o por algún desgraciado accidente que acababa con él hecho pedazos. Supongo que la fuente podía romperse también, pero para eso ya había que ser muy bestia, así que el refrán no contempla esa posibilidad.

En la actualidad empleamos esta frase para referirnos al hecho de que si repetimos una acción con demasiada insistencia podemos acabar teniendo algún contratiempo.

Por ejemplo, hagamos un paralelismo entre ir a la fuente con el cántaro y hacer un contraste de hipótesis. ¿Creéis que no tienen nada que ver?. Pues lo tienen: si hacemos contrastes de hipótesis de forma insistente podemos acabar llevándonos un disgusto, que no será otro que el de cometer un error de tipo I. Me explicaré para que no penséis que me he dado con el cántaro en la cabeza en uno de los viajes a la fuente.

Recordemos que siempre que hacemos un contraste de hipótesis establecemos una hipótesis nula (H_0) que dice que la diferencia observada entre los grupos de comparación se debe al azar. A continuación, calculamos la probabilidad de que la diferencia se deba al azar y, si es menor que un valor determinado (habitualmente 0,05), rechazamos H_0 y afirmamos que es altamente improbable que la diferencia se deba al azar, por lo que la consideramos real. Pero claro, altamente improbable no significa seguro. Siempre hay un 5% de probabilidad de que, siendo H_0 cierta, la rechazemos, dando por bueno un efecto que en realidad no existe. Esto es lo que se llama cometer un error de tipo I.

Si hacemos múltiples comparaciones la probabilidad de cometer un error aumenta. Por ejemplo, si hacemos 100 comparaciones, esperaremos equivocarnos aproximadamente cinco veces, ya que la probabilidad de equivocarnos en cada ocasión será del 5% (y la de acertar del 95%).

Así que podemos preguntarnos, si hacemos n comparaciones, ¿cuál es la probabilidad de tener al menos un falso positivo?. Esto es un poco laborioso de calcular, porque habría que calcular la probabilidad de 1,2,..., $n-1$ y n falsos positivos utilizando probabilidad binomial. Así que recurrimos a un truco muy utilizado en el cálculo de probabilidades, que es

calcular la probabilidad del suceso complementario. Me explico. La probabilidad de algún falso positivo más la probabilidad de ninguno será de 1 (100%). Luego la probabilidad de algún falso positivo será igual a 1 menos la probabilidad de ninguno.

¿Y cuál es la probabilidad de ninguno?. La de no cometer error en cada contraste ya hemos dicho que es de 0,95. La de no cometer errores en n contrastes será de $0,95^n$. Así que la probabilidad de tener al menos un falso positivo será de $1 - 0,95^n$.

Imaginaos que hacemos 20 comparaciones. La probabilidad de cometer, como mínimo, un error de tipo I será de $1 - 0,95^{20} = 0,64$. Habrá un 64% de probabilidad de que cometamos un error y demos por existente un efecto que en realidad no existe por puro azar.

Pues que chorrada, me diréis. ¿Quién se va a poner a hacer tantas comparaciones sabiendo el peligro que tiene?. Pues, si os paráis a pensarlo, lo habéis visto muchas veces. ¿Quién no ha leído un artículo sobre un ensayo que incluía un estudio post hoc con múltiples comparaciones?. Es bastante frecuente cuando el ensayo no da resultados con significación estadística. Los autores tienden a exprimir y torturar los datos hasta que encuentran un resultado satisfactorio.

Sin embargo, desconfiad siempre de los estudios post hoc. El ensayo debe tratar de responder a una pregunta previamente establecida y no buscar respuestas a preguntas que nos podemos hacer después de finalizarlo, dividiendo los participantes en grupos según características que no tienen nada que ver con la aleatorización inicial.

De todas formas, como es una costumbre difícil de erradicar, sí que podemos exigir a los autores de los ensayos que tengan una serie de precauciones si quieren hacer estudios post hoc con múltiples contrastes de hipótesis. Lo primero, todo análisis que se haga con los resultados del ensayo debe especificarse cuando se planifica el ensayo y no una vez terminado. Segundo, los grupos deben tener cierta plausibilidad biológica. Tercero, debe evitarse hacer comparaciones múltiples con subgrupos si los resultados generales del ensayo no son significativos. Y, por último, utilizar siempre alguna técnica que permita mantener la probabilidad de error de tipo I por debajo del 5%, como la [corrección de Bonferroni](#) o cualquier otra.

A nosotros nos quedará un último consejo: valorar con precaución las diferencias que se puedan encontrar entre los distintos subgrupos, sobre todo cuando los valores de p son discretos, entre 0,01 y 0,05.

Y aquí dejamos los estudios post hoc y sus trampas. No hemos comentado que hay más ejemplos de comparaciones múltiples además del análisis de

subgrupos postaleatorización. Se me ocurre el ejemplo de los estudios de cohortes que estudian diferentes efectos producto de una misma exposición, o el de los análisis intermedios que se hacen durante los ensayos secuenciales para ver si se cumple la regla de finalización preestablecida. Pero esa es otra historia...
