ReaR
ISNN 1989 4090
Revista electrónica de AnestesiaR
Febrero 2025

ORIGINAL

Percy Fawcett y la Ciudad Perdida. Tasa de descubrimiento falso.

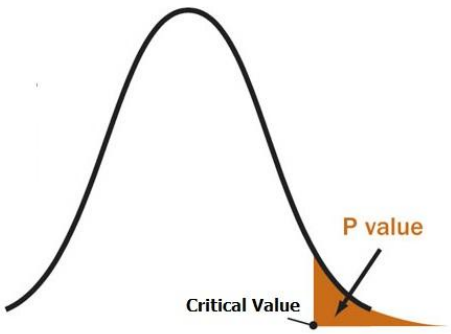
Molina Arias M

Hospital Infantil Universitario La Paz. Madrid. España.

Resumen

Cuando se realizan múltiples contrastes de hipótesis, se incrementa la probabilidad de cometer un error de tipo 1, aumentando el riesgo de detectar efectos falsamente positivos. La tasa de falso descubrimiento permite limitar la probabilidad de error de tipo 1 cuando el número de contrastes es muy elevado, permitiendo controlar también el riesgo de cometer errores de tipo 2 y fracasar en la detección de verdaderos positivos.

Introducción



Cuando se realizan múltiples contrastes de hipótesis, se incrementa la probabilidad de cometer un error de tipo 1, aumentando el riesgo de detectar efectos falsamente positivos. La tasa de falso descubrimiento permite limitar la probabilidad de error de tipo 1 cuando el número de contrastes es muy elevado, permitiendo controlar también el riesgo de cometer errores de tipo 2 y fracasar en la detección de verdaderos positivos.

Pocos conoceréis a Percival Harrison Fawcett, Percy para los amigos. Y aunque él sea un desconocido para el gran público, no ocurre lo mismo con el personaje que se creó inspirándose en su historia. Un tal Indiana Jones.

Percy fue un explorador inglés que, a principios del siglo XIX, se embarcó en una serie de expediciones al corazón del Amazonas en busca de una legendaria ciudad conocida como la Ciudad Perdida de Z, que él creía que era El Dorado, la legendaria ciudad que debía existir en la selva inexplorada de Brasil.

Armado con mapas antiguos, relatos de los indígenas y su férrea determinación, Fawcett y su equipo atravesaron ríos, selvas y enfrentaron numerosos peligros. Sin embargo, la selva amazónica estaba llena de “señuelos”: fragmentos de cerámica, restos de aldeas abandonadas y formaciones naturales que a menudo parecían indicios de la Ciudad Perdida. Cada vez que Fawcett encontraba uno de estos posibles indicios, su corazón se llenaba de esperanza, solo para enfrentarse a la desilusión cuando la evidencia resultaba ser una falsa pista.

No sabemos si Fawcett llegó a descubrir la Ciudad Perdida, ya que desapareció en la selva junto a su hijo Jack y un amigo de este, Raleigh Rimmell, el 29 de mayo de 1925, pero sí podemos estar seguros de su persistencia en seguir cada pista, inmune a la fiebre y la fatiga, a pesar de los numerosos falsos

positivos y engaños con los que la selva trataba de impedir su objetivo.

A mi esta actitud me parece similar al desafío que enfrentan los científicos con las comparaciones múltiples en sus investigaciones, sobre todo en aquellos campos en los que el *big data* y las bases de datos con miles de variables campan a sus anchas, como en las modernas ciencias "ómicas".

En esta selva estadística, cada análisis puede parecer una pista prometedora, pero sin métodos adecuados para controlar los errores, los investigadores pueden terminar siguiendo muchas pistas falsas. Si Fawcett hubiera tenido un método para discernir mejor cuáles pistas valía la pena seguir, podría haber optimizado su búsqueda y quizás haber encontrado su ciudad perdida.

De una forma parecida, los investigadores utilizan la tasa de descubrimiento falso para equilibrar el riesgo de falsos positivos y maximizar la probabilidad de hallazgos significativos. Seguid leyendo esta entrada y veremos cómo podemos mantener a raya los falsos positivos, siempre inevitables y casi obligados cuando se realizan numerosas comparaciones múltiples.

El planteamiento del problema

Ya vimos en una entrada anterior cómo el error de tipo 1 es uno de nuestros enemigos a batir en el campo de batalla de la investigación.

Cuando hacemos una comparación entre dos grupos, planteamos un contraste de hipótesis en el que establecemos una hipótesis nula que, en general, dice lo contrario de lo que queremos demostrar. La mayor parte de los casos se supondrá, bajo la hipótesis nula, que las diferencias observadas se deben únicamente al azar.

De esta forma, calculamos el valor de p, que no es otra cosa que encontrar una diferencia como la encontrada o mayor por causa del azar, siendo cierta la hipótesis nula. Solemos establecer el umbral de 0,05, de forma que si $p < 0,05$ consideramos poco probable que la diferencia se deba al azar y rechazamos la hipótesis nula: con ello "demostramos" que la diferencia se debe a la intervención o exposición en estudio.

Veis que he escrito "demostramos" entre comillas, y esto es porque nunca podemos estar seguros de tomar la decisión correcta, ya que hay una probabilidad del 5% (0,05) de que la hipótesis nula que estamos rechazando como falsa sea, en realidad, cierta. Estaremos detectando un efecto que no existe, un falso positivo. Este es el error de tipo 1.

Pensadlo un momento, en 1 de cada 20 contrastes de hipótesis en los que se rechaza la hipótesis nula metemos la pata y detectamos un falso positivo.

Pues bien, esta cifra, que a algunos les puede parecer pequeña, va creciendo cuando realizamos varios contrastes, como vimos en una entrada anterior.

Cuando realizamos múltiples comparaciones en un mismo estudio, la probabilidad de error de tipo 1 (la llamada alfa) en cada contraste individual es de 0,05, pero el alfa global aumenta con el número de comparaciones. Podemos calcular el alfa global, que es la probabilidad de obtener, al menos, un falso positivo, como $1-0,95^n$, siendo n el número de comparaciones.

Si hacemos un contraste de hipótesis, la probabilidad de obtener un falso positivo es de 0,05 (5%). Pero si realizamos 10 contrastes en el mismo estudio, esta probabilidad sube a 0,4

(40%): si repetimos el estudio muchas veces, en casi 1 de cada 2 cometeremos, al menos, un falso positivo.

Imaginad en un estudio de asociación genómica, en los que se estudia la asociación de determinado rasgo con, por ejemplo, 100 genes. En este caso, la probabilidad de obtener, al menos, un falso positivo es de 0,99 (99%). Si os parece mucho, pensad que habitualmente este tipo de estudios realizan miles o decenas de miles de comparaciones.

No hay duda, tenemos que controlar los falsos positivos.

La solución simple, pero insuficiente

Ya vimos en su momento que existen varios métodos para ajustar el umbral de p para considerar cada contraste significativo y mantener el alfa global por debajo de 0,05. En estos casos, por muchas comparaciones que hagamos, la probabilidad global de obtener, al menos, un falso positivo se mantiene por debajo de 0,05.

El método más sencillo y uno de los más utilizados es el de Bonferroni, que consiste en dividir el valor habitual de p (que suele ser 0,05) entre el número de contrastes o comparaciones, obteniendo así el nuevo umbral de significación estadística para cada contraste.

En el ejemplo de los 10 contrastes, el nuevo valor de p sería de 0,05 / 10 = 0,005. Esto quiere decir que, para rechazar la hipótesis nula en cada contraste individual, tenemos que obtener un valor de $p < 0,005$.

A nivel global este método es una buena solución. Si calculamos la probabilidad de obtener, al menos, un falso positivo con la fórmula anterior (sustituyendo 0,95 por 0,995), obtenemos un valor de $1 – 0,995^{10} = 0,0488$. Vemos que el alfa global se mantiene por debajo del valor elegido por convenio de 0,05.

Ahora pensemos qué falla con este método. El valor de p para rechazar la hipótesis nula es 10 veces más bajo. Esto quiere decir que aumentará la probabilidad de cometer un error de tipo 2: no alcanzar significación estadística y no poder rechazar la hipótesis nula cuando, en realidad, es falsa. Dicho de forma sencilla, aumenta el riesgo de no detectar efectos que en realidad existen y atribuir la diferencia falsamente al azar.

Si esto ocurre con 10 comparaciones, imaginad el ejemplo de los 100 genes. Aquí, la nueva p = 0,05 / 100 = 0,0005. Será cada vez más difícil detectar los verdaderos positivos con un valor de p tan bajo. Habremos conseguido mantener a raya los falsos positivos, pero a un precio demasiado alto: no poder detectar los verdaderos positivos.

Aquí es donde entra en juego una forma diferente de enfocar el tema: la tasa de descubrimiento falso.

Tasa de descubrimiento falso

Ya hemos visto que el método de Bonferroni es demasiado restrictivo cuando el número de comparaciones es alto.

Sabemos que, cuando hagamos muchas comparaciones, de entre todos los contrastes en los que rechacemos la hipótesis nula habrá un número variable de falsos positivos.

De esta manera, podemos plantearnos un método que nos asegure que la proporción de falsos positivos no supere determinado valor respecto al número de rechazos totales, ya sea un 10%, 20% o el valor que consideremos más conveniente.

Se define así la mal llamada tasa de descubrimiento falso (FDR, por sus siglas en inglés) como la expectativa de la proporción de falsos descubrimientos (rechazos incorrectos de la hipótesis nula) sobre el total de descubrimientos (rechazos de hipótesis nulas).

Si llamamos V al número de falsos positivos y R al número total de rechazos, esta proporción (en realidad no es una tasa) puede expresarse según la siguiente fórmula:

FDR = E(V / R)

E representa la media esperada para esta proporción. ¿Qué quiere decir esperada? Imaginemos que repetimos el estudio infinidad de veces. Si establecemos una FDR de 20%, obtendremos, de media, un 20% de falsos positivos. En unos estudios serán más y en otros serán menos. El azar, nuestro inseparable compañero.

Este enfoque, útil cuando se realice un alto número de comparaciones, permite una mayor flexibilidad y poder estadístico en la detección de efectos verdaderos, evitando el conservadurismo extremo de otros métodos como el de Bonferroni. En esencia, la FDR permite a los investigadores manejar el balance entre descubrir efectos reales y minimizar falsos positivos de manera más eficiente y pragmática.

Vamos a ver cómo podemos establecer el nuevo valor de p para mantener la FDR en el valor que elijamos ya que, a diferencia del valor de p, no existe un valor general consensuado y dependerá de los datos y del contexto del estudio.

El método de Benjamini–Hochberg

El método de Benjamini-Hochberg es una de las alternativas sencillas para el cálculo del valor de p que nos permita controlar la FDR deseada. Para realizarlo, este algoritmo sigue una serie de pasos secuenciales:

1. Especificar el valor máximo de descubrimientos falsos que deseamos al realizar las comparaciones múltiples. Vamos a llamar q a este valor que controla la FDR.

2. Calcular los valores de p para los m contrastes de hipótesis que queramos realizar.

3. Ordenar los m valores de p de menor a mayor.

4. Seleccionamos, de entre los valores de p, aquellos que son menores que el resultado de multiplicar q por el número de orden de la p y dividirlo por el total de contrastes (m).

5. Este valor será el nuevo valor de significación estadística. El valor de p de un contraste determinado deberá ser inferior para rechazar la hipótesis nula.

Por si no se ha entendido bien, vamos a ver un ejemplo sencillo. Voy a poner solo 5 comparaciones, aunque ya sabemos que, en la práctica, suelen hacerse muchas más.

Supongamos que los valores de p de los 5 contrastes son 0,001, 0,583, 0,123, 0,012 y 0,473. Queremos un FDR = 10% (0,1).

Primero, ordenamos los valores de p de menor a mayor: $p_1 = 0{,}001$, $p_2 = 0{,}012$, $p_3 = 0{,}123$, $p_4 = 0{,}473$ y $p_5 = 0{,}583$.

Ahora realizamos el cuarto paso del algoritmo: $p_1 = 0{,}001 < 0{,}1 \times (1/5)$, $p_2 = 0{,}012 < 0{,}1 \times (2/5)$, $p_3 = 0{,}123 > 0{,}1 \times (3/5)$, $p_4 = 0{,}473 > 0{,}1 \times (4/5)$ y $p_5 = 0{,}583 > 0{,}1 \times (5/5)$.

Solo cumplen la condición p_1 y p_2, siendo esta última la que tiene el valor máximo de las dos, 0,012. Resultado: si utilizamos este valor como el nuevo umbral de significación estadística (en lugar de 0,05), por término medio, el

10% de las veces que rechacemos la hipótesis nula incurriremos en un error de tipo 1 o, lo que es lo mismo, detectaremos un falso positivo.

Aunque este valor pueda parecer alto, pensad que suele emplearse en estudios de carácter exploratorio, como la asociación con miles de genes, que se utilizan para abrir nuevas vías de investigación. Habrá que confirmar los hallazgos encontrados.

Para ir acabando, quiero que os fijéis en una diferencia entre los dos métodos de ajuste que hemos comentado. En el método de Bonferroni, el nuevo valor de p no depende de los datos, sino únicamente del número de contrastes que queramos realizar, por lo que podemos conocer de antemano el valor ajustado de p.

Sin embargo, esto no es posible con el método de Benjamini-Hochberg, ya que el nuevo valor de p dependerá de los datos y de los valores de p de los diferentes contrastes. Es algo similar a lo que ocurre con el método de Holm, otro método de ajuste para comparaciones múltiples que se utiliza en condiciones similares a las del método de Bonferroni.

Nos vamos…

Y aquí lo vamos a dejar por hoy.

Hemos visto cómo el azar nos acompaña permanentemente en todos nuestros experimentos y cómo parece una regla de oro el hecho de que siempre tengamos que renunciar a algo para poder mejorar otra cosa diferente. Un poco como la vida misma.

Nos ha quedado claro cómo el riesgo de cometer un error de tipo 1 y detectar un falso positivo aumenta con el número de comparaciones que realizamos, por lo que se hace necesario aplicar algún método de ajuste.

Cuando son pocas comparaciones, podemos recurrir al ajuste del valor de p mediante métodos como el de Bonferroni, con los que mantendremos a raya los falsos positivos perjudicando poco los verdaderos positivos.

Pero cuando hay muchas comparaciones, estos métodos pueden impedirnos detectar efectos reales, por lo que es necesario cambiar el tipo de enfoque y resignarse a tener cierta proporción de falsos positivos para poder seguir detectando los verdaderos.

Hemos visto el método de Benjamini-Hochberg para ajustar nuestra tasa de descubrimientos falsos, aunque no es el único método posible. Para los virtuosos, existen diversos métodos que emplean técnicas de remuestreo para el cálculo de p y el ajuste de la FDR. Pero esa es otra historia…

Bibliografía

– Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, et al. A practical guide to methods controlling false discoveries in computational biology. Genome Biol. 2019;20:118. (PubMed)

– Multiple testing. En: James G, Witten D, Hastie T, Tibshirani R, Taylor J, eds. An introduction to statistical learning with applications in Python. Springer International Publishing AG, 2023: 557-96. (HTML)

Correspondencia al autor

Manuel Molina Arias
mma1961@gmail.com
Servicio de Gastroenterología.
Hospital Infantil Universitario La Paz.
Madrid. España.

Aceptado para el blog en mayo de 2024