Rev Chil Salud Pública 2012;
Vol 16 (3): 225 -232

Artículo original

Ajuste del valor-p por contrastes múltiples

RESUMEN

Objetivo: *Este trabajo tiene un objetivo doble: Por un lado, se plantea la discusión de cuándo se debe ajustar el p-valor por contrastes múltiples. Por otro, una vez decidido si se debe realizar este ajuste, se revisan algunas de las alternativas propuestas al método de Bonferroni.* **Estudio de simulación***: Las distintas posibilidades que se presentan son estudiadas a partir de un pequeño estudio de simulación que ilustra las limitaciones y las bondades de los distintos procedimientos considerados.* **Conclusiones***: La consecuencia directa del avance tecnológico ha sido el aumento en la cantidad de información disponible. El análisis de esa información, usualmente basado en herramientas estadísticas, debe ser manejado con extremo cuidado en aras de evitar malinterpretaciones y descubrimientos espurios.*

Palabras clave: *Contrastes múltiples, Ajuste del p-valor, Family wise error rate, false discovery rate.*

P-VALUE ADJUSTMENT FOR MULTIPLE COMPARISONS

ABSTRACT

Objective: *This study has double objectives: on one hand, it discusses when the p-value should be adjusted for multiple comparisons. On the other, some alternatives to the Bonferroni method are discussed, for when adjustment has been decided.* **Simulation study:** *The different possibilities that are presented are based on a small simulation study, which illustrates the limitations and benefits of the different procedures being considered.* **Conclusions:** *The direct consequence of technological advancement has been the increase in the quantity of available information. The analysis of that information, usually with statistical tools, should be managed with extreme caution in order to avoid misinterpretations and deceptive discoveries.*

Key words: *Multiple comparisons, p-value adjustment, family wise error rate, false discovery rate.*

PABLO MARTÍNEZ-CAMBLOR [(1)]

[(1)]Oficina de Investigación Biosanitaria del Principado de Asturias C/ Matemático Pedrayes, 25, Entresuelo 33005 Oviedo. Asturias. España.

Este trabajo ha recibido financiación del proyecto MTM2011-23204 del Ministerio de Ciencia e Innovación del Gobierno de España (Fondos FEDER incluidos).

Introducción

El desarrollo tecnológico facilita la recopilación de gran cantidad de información, también la aparición de herramientas que permiten un rápido análisis de la misma. En particular, en los estudios biomédicos se suele tener la capacidad de recoger mucha información procedente de diferentes fuentes: historias clínicas, marcadores biológicos, información genética, etc. Paralelamente, en una misma investigación empírica se intenta responder a distintas preguntas. Dado que la veracidad de las hipótesis planteadas suele someterse a un juicio estadístico cuya herramienta fundamental es el cálculo de probabilidades, tratar estas hipótesis individualmente puede aumentar la probabilidad global de error de Tipo I (falsos positivos) dando lugar a interpretaciones erróneas de las conclusiones obtenidas[1]. Por este motivo, en ocasiones es necesario realizar algún tipo de corrección en la significación final obtenida. Sin embargo, el uso de estos ajustes en biomedicina ha suscitado y suscita cierta controversia.

Una opinión ampliamente extendida entre los epidemiólogos considera que el ajuste por contrastes múltiples es, no solamente innecesario, sino que entorpece la correcta interpretación de la inferencia estadística[2,3] aumentando, de forma innecesaria, la probabilidad de error de Tipo II (falsos negativos). Desde un punto de vista técnico, más estadístico, se refutan los argumentos anteriores y se sigue recomendando tener mayor cuidado con el manejo e interpretación de los p-valores obtenidos[4,5].

En este trabajo se plantea un objetivo doble: Por un lado, ahondar en el problema de ajustar o no el valor-p por el número de contrastes realizado. Sobre nuestra visión en este aspecto versa el segundo apartado. Por el otro, una vez que se ha decidido corregir las significaciones obtenidas por el número de contrastes realizados, se ofrece una aproximación técnica (aunque, con la intención de no ser excesivamente técnica y evitando, en lo posible, ambages matemáticos) a este problema. Revisamos algunas de las soluciones clásicas ofrecidas, las cuales son usualmente menos conservadoras que el clásico método de Bonferroni. En el cuarto apartado se describe un pequeño estudio de simulación en el que se analizan los métodos descritos en la tercera sección. Para finalizar, en el quinto apartado, exponemos nuestras conclusiones.

¿Ajustar o no ajustar? esa es la cuestion

Convencionalmente, para contrastar la veracidad de una determinada hipótesis (nula), se fija un nivel de significación (probabilidad de error de Tipo I que se está dispuesto a asumir, usualmente, el bien conocido $\alpha = 0.05$), se elige un test estadístico y se obtiene un valor-p (este puede interpretarse como el apoyo o la credibilidad que tiene la hipótesis nula). La hipótesis nula se rechaza si el p-valor es menor que el nivel de significación fijado y no se rechaza en caso contrario. El problema surge cuando el número de hipótesis a contrastar (o el número de pruebas estadísticas realizados) es más de uno, esto es, se tienen varios p-valores.

Técnicamente, si se contrastan *k*-tests a un nivel de significación α, la probabilidad de que haya al menos un falso positivo es mayor de α (en concreto, asumiendo independencia entre las k hipótesis contrastadas, la probabilidad es de $1-(1-\alpha)^k$ (ver Tabla 1) y se debería utilizar algún método de ajuste que garantice que se está respetando el error de Tipo I fijado previamente. En la Tabla 1 se muestra la probabilidad de que exista al menos un falso positivo en función del número de pruebas estadísticas realizadas.

Tabla 1. Falsos positivos esperados y probabilidad de error de Tipo I (al menos haya un falso positivo) en función del número de hipótesis nulas independientes contrastadas.

Número de hipótesis independientes	Falsos positivos esperados	Probabilidad de error de Tipo I (α=0.05)
1	0	0.050
20	1	0.641
40	2	0.871
100	5	0.994

Tabla 2. Esquema de los errores que se plantean cuando se contrastan, simultáneamente, k hipótesis nulas[9].

	Hipótesis declaradas verdaderas	Hipótesis declaradas falsas	
Nº de hipótesis verdaderas	U	V	k0
Nº de hipótesis falsas	T	S	k-k0
	k-R	R	k

El principal argumento esgrimido por los detractores del uso de métodos de ajuste por contrastes múltiples es que la hipótesis nula que se contrasta al realizar ajustes por multiplicidad es la intersección de todas las hipótesis nulas involucradas en el estudio (esto es, que todas las hipótesis nulas sean ciertas) y que, en realidad, esta hipótesis suele tener poco interés en la investigación[6]. Además, el ajuste por contrastes múltiples implica que un determinado resultado (con su p-valor asociado) debe ser interpretado de forma diferente en función del número de tests que hayan sido realizados simultáneamente en ese estudio[2]. Inmediatamente, surge la pregunta; ¿qué tests deben ser tenidos en cuenta a la hora de realizar el ajuste? La respuesta no es trivial ya que en este conjunto podrían estar todos los tests incluidos en el trabajo en cuestión (léase artículo, reporte técnico u otro documento), todos los tests realizados en ese estudio, hayan sido considerados en el mismo trabajo, en distintos trabajos o, incluso, desechados. Todos los tests realizados en trabajos con objetivos similares. Obviamente, las implicaciones serían, cuanto menos, extrañas. Un valor-p podría verse afectado por estudios futuros (con objetivos similares) y las conclusiones obtenidas estar constantemente sujetas a posibles cambios.

Desde un punto de vista más técnico (estadístico), se sostiene que, en realidad, los procedimientos de ajuste por multiplicidad no solo atañen a la hipótesis intersección, sino que posibilitan alcanzar una conclusión más real para todas las hipótesis involucradas en el estudio, permitiendo deducir qué hipótesis se rechazan y cuáles no, manteniendo, además, el adecuado control sobre el error de Tipo I[4]. Asimismo, se destaca que, en estudios confirmatorios, en los cuales los objetivos están prefijados y representados como contrastes múltiples y, en los que los tests de significación son usados como una herramienta estadística para la evaluación de la decisión final a tomar, la realización de ajustes no solamente es recomendable sino que es obligatoria[7].

Obviamente, nuestra posición es más cercana al segundo planteamiento que al primero. Compartimos con Bender y Lange[4] la idea de que, en estudios exploratorios, aunque las hipótesis involucradas no (suelen) están predefinidas, no es necesario ajustar los p-valores por el número de tests realizados (notar que este número puede ser muy elevado aunque la información recogida no sea excesiva ya que se incrementa con el número de variables, número de subgrupos (edad, sexo,...), posibles segmentaciones de la base de datos, etc.), si bien, tanto el investigador como los lectores deben tener claro que el estudio es exploratorio y que todas las conclusiones deben ser tomadas con prudencia a la espera de su confirmación. A nuestro entender, el principal *gap* entre las dos visiones del problema radica en la negrita de la frase:

"...la probabilidad de que haya ***al menos un*** *falso positivo es mayor de α."*

Si se tiene un estudio, con una o varias hipótesis nulas, se realizan los correspondientes tests estadísticos y se derivan las conclusiones relativas a los p-valores obtenidos teniendo en cuenta el nivel de significación fijado previamente, se sabe que la posibilidad de que cada una de las hipótesis planteadas sea un falso positivo es del 5% (asumamos por un momento que el nivel de significación α es el habitual 0.05). Lógicamente, la probabilidad de que *al menos una* de esas hipótesis planteadas sea un falso positivo es mayor de 0.05, y así debe entenderlo tanto la/el investigador/a como las/los lectoras/

Tabla 3. Porcentaje de aciertos (las conclusiones sobre todas las hipótesis son ciertas) observados en 2.000 réplicas de Monte Carlo[a].

Nº de Hipótesis Falsas	% Medio de acierto									
	r = 0					r = 3/4				
	C	B	H	O	W	C	B	H	O	W
100	93.3	43.0	52.1	64.5	45.3	93.1	42.5	55.3	72.7	66.1
50	94.1	71.5	73.1	73.8	72.2	94.2	71.9	73.6	74.8	81.7
25	94.6	85.7	86.0	86.1	86.0	94.6	85.9	86.2	86.3	90.5
10	95.0	94.3	94.3	94.3	94.4	95.6	94.4	94.6	94.6	96.4
0	95.0	99.9	99.9	99.9	99.9	94.7	99.9	99.9	99.8	99.9

[a] Cuando no se realiza ninguna corrección (C), se utiliza el método de Bonferroni (B), el de Holm (H), el de Hommel (O) y el de Westfal y Young (W) (basado en 1 000 permutaciones). El número total de hipótesis consideradas es 100 (n=25).

es. El problema surge cuando la investigación se desorienta y se asume que el valor-p más bajo (o los valores-p por debajo del nivel de significación) es el resultado principal de la misma, sin tener en cuenta que, la probabilidad de que, bajo la hipótesis nula, ese valor esté por debajo del nivel de significación fijado (α) es superior a α. Insistimos, en concreto, en el caso de hipótesis independientes, de $1-(1-\alpha)^k$ (valor cercano a 1 para 100 contrastes independientes, Tabla 1). Este error se agrava en determinados contextos. En particular, en investigaciones que involucran marcadores genéticos, se pueden considerar varias decenas (incluso centenas y hasta millares) de variables y, en muchas ocasiones, directamente se buscan aquellas cuyo valor-p está por debajo del nivel de significación. Obviamente, en estos casos, es necesario ajustar, ya que para que el punto de corte para el valor-p respete la probabilidad de error de Tipo I fijado, este debe tener en cuenta si se elige entre el mínimo (o los más pequeños) de una serie de p-valores y no uno (perteneciente a una hipótesis nula particular) concreto. Otro caso especialmente delicado ocurre cuando existen distintos posibles tests para contrastar una determinada hipótesis (contrastes paramétricos, no paramétricos, etc.), se calculan todos y se elige el más conveniente para las conclusiones deseadas. Un ejemplo claro pueden ser las curvas de supervivencia: Muchos de los paquetes estadísticos habituales incorporan varios tests para contrastar la igualdad de curvas de supervivencia; la principal diferencia entre ellos radica en el tipo de diferencias entre las curvas que detectan[8] (al principio de la curva, al final, diferencias proporcionales, etc.). La práctica adecuada (si se quiere mantener la probabilidad de error de Tipo I) es pre-fijar el test a utilizar antes de ver donde están las diferencias en las curvas involucradas o bien, ajustar los p-valores por el número de tests realizados. De no hacerlo así, el error de Tipo I que se cometerá será superior al fijado en el estudio.

Bonferroni y sus alternativas

Una vez que se ha decidido ajustar los p-valores por el número de tests realizados, la pregunta pasa a ser cómo hacerlo.

Aunque se han desarrollado numerosos métodos multivariantes con el objetivo de realizar contrastes múltiples, el método de Bonferroni (utiliza como punto de corte α/k, donde alfa es el nivel de significación y k el número de contrastes o, equivalentemente, se pasa a trabajar con los p-valores transformados: $P^*_i = min(1, k \cdot P_i),\ 1 \le i \le k$) es el más conocido y principalmente utilizado, permitiendo contrastar de forma sencilla y sin asunciones adicionales cada una de las hipótesis (nulas) individuales involucradas en el estudio. Sin embargo, este método es muy conservador y aumenta considerablemente la probabilidad de error de Tipo II. En este apartado se describen algunos métodos adicionales.

Consideremos un problema en el que se pretenden contrastar simultáneamente k hipótesis (nulas), H_{0_i} ($1 \le i \le k$), de las cuales k_0 son ciertas (k-k_0, falsas). Supongamos que para un nivel de significación prefijado α, R (=R(α)) hipótesis son rechazadas. La Tabla 2 resume esta información de manera estándar (esta tabla es similar a la Tabla 1 que aparece en[9]).

Lógicamente, k (número total de hipótesis nulas involucradas) y R (número de hipótesis recha-

zadas) son valores conocidos mientras que U (número de hipótesis nulas verdaderas no rechazadas), V (número de hipótesis nulas verdaderas rechazadas, falsos positivos), S (número de hipótesis nulas falsas declaradas falsas, verdaderos positivos) y T (número de hipótesis falsas declaradas verdaderas, falsos negativos) son no observables y, por lo tanto, desconocidos por el investigador. Convencionalmente, el objetivo de los métodos de ajuste por contrastes múltiples es controlar la probabilidad de cometer algún error en la familia de contrastes considerados, conocida por su acrónimo en inglés, FWER (Family Wise Error Rate), esto es, controlar $P(V \geq 1)$. Notar que, contrastando cada hipótesis individual a nivel α/k se tiene que:

$$P(V \geq 1) \leq \sum_{i=1}^{k} P_{H_{0,i}}(\text{Rechazar } H_{0,i}) = \sum_{i=1}^{k} \alpha/\kappa = \alpha.$$

Es importante reseñar que la igualdad solamente se consigue cuando las hipótesis a contrastar son independientes, dado que esta hipótesis es poco habitual, usualmente, el método de Bonferroni es excesivamente conservador.

Varios autores han propuesto procedimientos que, tratando de mantener la sencillez y la rapidez de cálculo de la corrección de Bonferroni, son menos conservadores y, por tanto, más potentes (como es sabido, la potencia de un test es el complementario de su probabilidad de error de Tipo II). Holm[10] propuso un método que, a pesar de admitir sensibles mejoras[11], sigue manteniendo ciertas ventajas sobre sus competidores que lo hacen más atractivo[12]. Si $H_{0,1}$, $H_{0,2}$,..., $H_{0,k}$, son las k hipótesis nulas que se desean contrastar y P_1, P_2,..., P_k, sus respectivos p-valores (crudos, sin ningún tipo de ajuste),

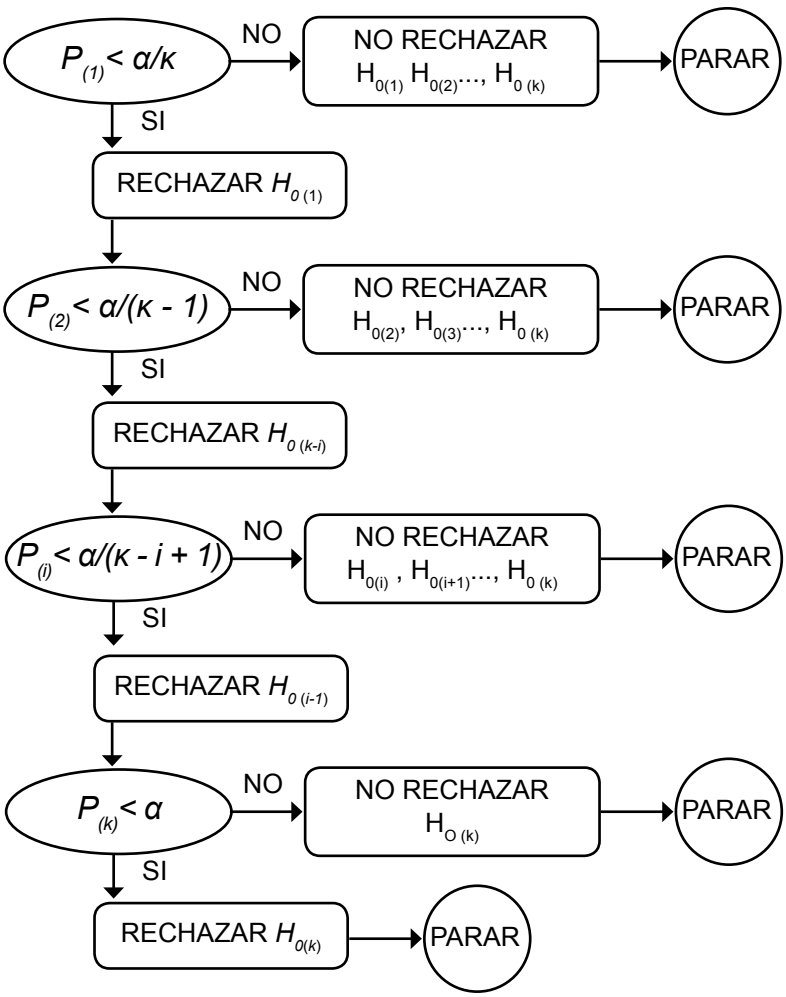

Figura 1. Representación gráfica del procedimiento (algoritmo) de ajuste descrito por Holm[10].

el procedimiento de ajuste propuesto por Holm es el siguiente:

A_1. Sean $P_{(1)}, P_{(2)},..., P_{(k)}$ los p-valores ordenados, esto es, $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(k)}$, y sean $H_{0,(1)}, H_{0,(2)},\cdots, H_{0,(k)}$ sus respectivas hipótesis nulas.

A_2. Calcular: j = *máximo (i ∈1,...,k, tal que* $(k - l + 1) \cdot P_{(l)} < \alpha$, *para todo* $1 \leq l \leq i$*).*

A_3. Se rechazan las hipótesis correspondientes a los p-valores: $P_{(1)},..., P_{(l)}$ y no se rechazan las hipótesis correspondientes a los p-valores $P_{(l+1)},..., P_{(k)}$. Esto es, se rechaza $H_{0,(1)},..., H_{0,(l)}$ y no se rechaza $H_{0,(l+1)},..., H_{0,(k)}$.

En la Figura 1 se representa, en forma de algoritmo, las diferentes iteraciones del procedimiento de Holm.

Este procedimiento supone una mejora substancial en la potencia obtenida respecto del método de Bonferroni. Sin embargo, Hommel[13] muestra que, para determinados problemas (dependiendo del número de hipótesis nulas que realmente son falsas) su probabilidad de error de Tipo I es superior a α. El primer paso del algoritmo propuesto por Hommel[13] coincide con A_1. El criterio de decisión final es ligeramente diferente al propuesto por Holm[10]:

O_1. Sean $P_{(1)}, P_{(2)},..., P_{(k)}$ los p-valores ordenados, esto es, $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(k)}$, y sean $H_{0,(1)}, H_{0,(2)},\cdots, H_{0,(k)}$ sus respectivas hipótesis nulas.

O_2. Calcular: j = *máximo (i ∈1,...,k, tal que* $P_{(k-i+l)} \geq l \cdot \alpha / i$, *para todo* $1 \leq l \leq i$*).*

O_3. Si j no existe, se rechazan todas las hipótesis nulas. Si existe, se rechazan aquellas hipótesis nulas tales que $P_i < \alpha / j$.

A pesar de la reducción en el error de Tipo II que estos procedimientos consiguen, cuando

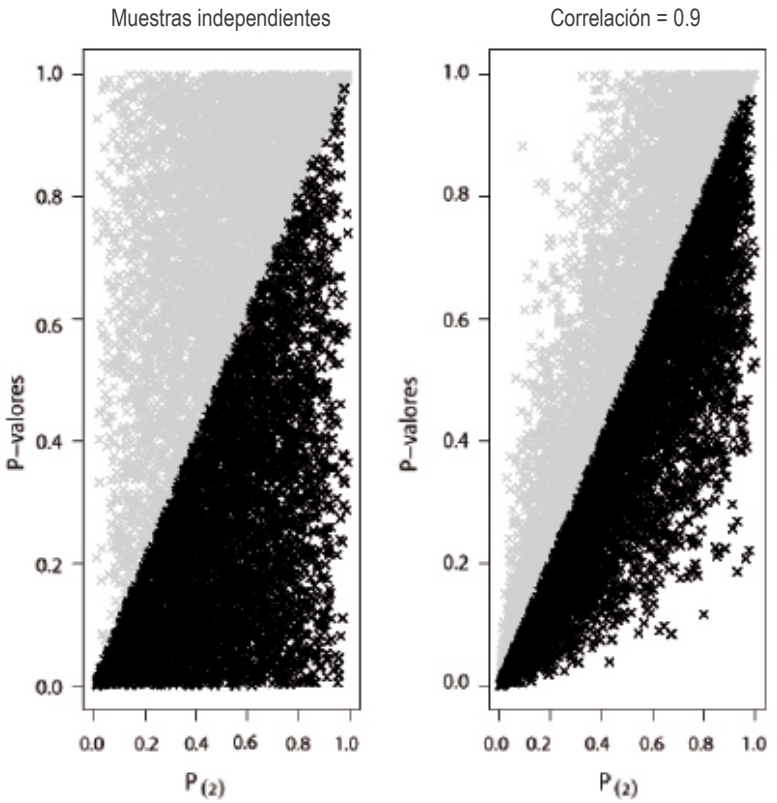

Figura 2. A la izquierda, p-valores obtenidos en una comparación de medias para dos muestras (n=25) generadas desde distribuciones normales independientes (prueba T de Student) para tres hipótesis nulas independientes. A la derecha, la correlación entre las muestras es de 0.9. Los puntos grises representan $P_{(3)}$ (p-valor máximo), los puntos negros $P_{(1)}$ (p-valor mínimo). Ambos gráficos basados en 5,000 repeticiones.

las hipótesis involucradas están fuertemente relacionadas, siguen siendo extremadamente conservadores. En la Figura 2 se muestra la distribución del mínimo de tres valores-p para muestras independientes (izquierda) y muy correlacionadas (derecha). Se puede apreciar la diferencia en la dispersión entre ambos modelos y el error que supone asumir independencia cuando realmente no la hay.

Lógicamente, cualquier método que no asuma independencia entre las hipótesis involucradas, debe estimar las relaciones entre estas con la información disponible, esto es, desde los propios datos. Westfall y Young[14] proponen corregir el mínimo valor-p por su distribución muestral. Esta distribución se estima mediante técnicas de remuestreo (en concreto, mediante el método de permutaciones). El algoritmo seguido es el siguiente:

W_1. Para cada $1 \leq i \leq k$, se remuestrea desde la hipótesis nula. Para ello, se permutan las etiquetas de pertenencia al grupo, creando un conjunto artificial de datos bajo la hipótesis nula. Con esos datos, se calculan los correspondientes p-valores: P^b_1, P^b_2,⋯, P^b_k, y se calcula su mínimo, m^b=*mínimo* $\{P^b_1, P^b_2, \cdots, P^b_k\}$.

W_2. El paso anterior se repite B-veces (B es un número elegido por el investigador, preferentemente grande)**.** El nuevo nivel de significación será el que ocupe el percentil 5 en $\{m^1, m^2, \cdots, m^B\}$.

El principal inconveniente de este método es que, debido al hecho de estar basado en permutaciones, el número de cálculos involucrados es elevado y, por tanto, es computacionalmente lento cuando el número de hipótesis es alto. Notar que, asumiendo independencia, el método es similar al de Bonferroni.

Pequeño estudio de simulación

Con el objetivo de ilustrar el funcionamiento de los métodos anteriormente descritos, se realizó un pequeño estudio de simulación. En el problema considerado, se tienen 100 variables normalmente distribuídas y recogidas en dos poblaciones independientes (el tamaño muestral de cada una de ellas es **n**=25). En una de ellas, las medias son siempre 0 y las desviaciones típicas 1. En la otra, bajo la hipótesis nula, lógicamente tienen los mismos parámetros y, bajo la alternativa, la media es 1.44 y la desviación típica 1 (la probabilidad de error de Tipo II en cada hipótesis particular es de 0.05). La correlación entre los pares de variables es ***r*** (fueron considerados los casos ***r*** = 0 y ***r*** = 3/4).

La Tabla 3 muestra el porcentaje de acierto observado en 2 000 réplicas de Monte Carlo cuando no se utiliza ningún tipo de corrección (C), se corrige por el método de Bonferroni (B), por el de Holm (H), por el de Hommel (O) y por el método de permutaciones de Westfall y Young (W), este último, con B = 1 000.

Lógicamente, el método crudo (sin ningún tipo de corección), comete el 5% de errores en todas las muestras, lo que conlleva un 5% de error global. Debemos recordar que, en el problema considerado, ambos errores (tipo I y II) tienen la misma probabilidad de aparición (0.05). El método de Holm y el de Hommel obtienen resultados similares. El método de Westfal obtiene mejores resultados que el de Bonferroni para muestras correlacionadas.

Conclusiones

El avance tecnológico ha provocado que el volumen de información disponible crezca dramáticamente. Esto hace que, en el análisis de esa información, las herramientas estadísticas utilizadas, basadas usualmente en el cálculo de probabilidades, deban ser manejadas con extremo cuidado en aras de evitar descubrimientos espurios que ponen en riesgo el prestigio de la ciencia y la credibilidad de los avances científicos[15, 16]. La lógica nos dice que, para encontrar un efecto estadísticamente significativo, es suficiente con aumentar suficientemente el número de variables estudiadas.

La opinión más generalizada y que nosotros compartimos es que el *quid* de la cuestión radica en describir exactamente lo que se hace y tomar las evidencias detectadas en estudios observacionales con la debida cautela sin olvidar que, al aumentar el número de pruebas realizadas, la probabilidad de encontrar significaciones bajas (por debajo del nivel de significación) aumenta. Las conclusiones en este tipo de estudios, no deben basarse exclusivamente en el conocido valor-p (el nivel de significación no marca la frontera entre el bien y el mal) y otros criterios deben ser tenidos en cuenta (magnitud

de las diferencias, evidencias previas, coherencia clínica, etc.). En ensayos clínicos, corregir el valor-p por el número de pruebas realizadas debe ser y es un requisito indispensable y exigido por las agencias reguladoras. Obviamente, el método de ajuste utilizado, debe ser incluido en el protocolo del estudio como parte de las herramientas estadísticas utilizadas.

Con la explosión de las ciencias -ómicas, el problema de la multiplicidad ha ganado en importancia (citando a Berger: "*..the science is choking on the multiplicity problem*...")[17] y son innumerables los métodos que, con este fin y diferentes enfoques han sido desarrollados[18-20]. Los métodos basados en el control FWER han ido perdiendo interés y otras filosofías como las basadas en la proporción de falsos positivos[9] o FDR (acrónimo del inglés de False Discovery Rate), en el número esperado de hipótesis nulas rechazadas, método conocido como SGoF[18] (Sequential Goodness of Fit) o, y cada vez con más fuerza, en interpretaciones bayesianas[20] centran el interés de bioestadísticos y bioinformáticos. El estudio en profundidad de estos métodos es complejo y queda fuera de los objetivos de este trabajo.

Finalmente, reseñar que los métodos descritos en este trabajo, son una muestra de métodos clásicos que, sin ser tan conservadores como el clásico Bonferroni, nos permiten ajustar, de alguna manera, los valores-p obtenidos de forma que, sin rebajar excesivamente la potencia de los estudios, se controla más eficientemente el error de Tipo I.

Agradecimientos

El autor desea mostrar su agradecimiento al Editor y a los Revisores cuyas sugerencias han mejorado de forma sustancial el presente documento.

Referencias

1. bland JM, Altman DG. Multiple significance tests: the Bonferroni method. British Medical Journal, 1995;310:170.
2. Perneger TV. What's wrong with Bonferroni adjustments. British Medical Journal, 1998;316:1236-1242.
3. Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. American Journal of Epidemiology 1995;142: 904–908
4. Bender R, Lange S. What's wrong with arguments against multiplicity adjustments, Letter to the editor concerning BMJ 1998;316:1236-1238).
5. Thompson JR. Invited Commentary: Re: Multiple comparisons and related issues in the interpretation of epidemiologic data. American Journal of Epidemiologic, 1998: 147, 801-806.
6. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology, 1990;1:43-46.
7. Sankoh AJ, Huque MH, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials, Statistics in Medicine, 1997, 16(22), 2529-2542.
8. Leton E, Zuluaga P. Cómo elegir el test adecuado para comparar curvas de supervivencia. Medicina Clínica, 2006;127(3);96-99.
9. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 1995:57(1); 289-300.
10. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 1979: 6(2);60-65.
11. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika, 1986: 73(3), 751-754.
12. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: Bonferroni vs. Holm methods. American Journal of Public Health, 1996: 86(5);726-728.
13. Hommel G. A stagewise rejective multiple test procedure based on modified Bonferroni test, Biometrika, 1988: 75(2), 383-386.
14. Westfall PH, Young SS. Resampling-based multiple testing, 1993. Wiley, New York.
15. Ioannidis JP, Why most published research findings are false. PLoS Med. 2005. 2(8): e124.
16. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008, 19(5), 640-8.
17. Berger JO, What are the open problems in Bayesian statistics? Int. Soc. Bayesian Anal. Bull., 2011: 18, 1–4.
18. de Uña-Alvarez J, Carvajal-Rodriguez A, 'SGoFicance Trace': assessing significance in high dimensional testing problems, PLoSOne, 2010: 5(12): e15930.
19. Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, Blunt I, Wood C, Grigg O, Statistical methods for healthcare regulation: rating, screening and surveillance, J. R. Statist. Soc. A, 2012, 175, 1–47.
20. Scott JG, Berger JO, Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem, Ann. Statist, 2010: 38(5). 2587-2619.

Recibido: 13 de abril de 2012
Aprobado: 29 de junio de 2012