Análisis de subgrupos y de objetivos secundarios .
El problema de las comparaciones múltiples.
Comparación de valores basales.

Preparado por Luis M. Molinero (Alce Ingeniería)
CorreoE: **bioestadistica@alceingenieria.net**

Artículo en formato PDF

Junio 2002
www.seh−lelha.org/stat1.htm

Planteamiento del problema

En los ensayos clínicos aleatorizados y en los estudios observacionales además de estimar un efecto global medio, puede ser interesante calcularlo en subgrupos concretos de pacientes, con el ánimo de conocer mejor los mecanismos de actuación del tratamiento en cuanto a qué pacientes con determinadas características obtendrán mayor beneficio, o en su caso conocer aquellos que tienen un mayor riesgo de presentar la enfermedad, el efecto adverso, etc. Por otro lado, en muchas ocasiones, el objetivo principal de un estudio se calcula, por sus propias características, mediante más de una variable resultado, como puede ser un trabajo en el que interesa determinar tanto la eficacia como la tolerancia de un nuevo fármaco, comparándolas con el tratamiento habitual, y donde tenemos entonces al menos dos variables resultado: eficacia y tolerancia. Además se suele distinguir entre uno o más objetivos principales, y una o más variables secundarias. También es frecuente que para evitar controversias, por lo que comentaremos en el artículo, algunos investigadores fijen una sola variable principal en su trabajo y el resto las consideren como secundarias.

En el razonamiento estadístico clásico de contraste de hipótesis se calcula −suponiendo cierta la hipótesis nula (igualdad entre los tratamientos)− cuál es la probabilidad de observar una diferencia como la que de hecho hemos obtenido o más extrema, y se rechaza esa hipótesis nula sólo si el valor de probabilidad es inferior a uno prefijado (habitualmente 0.05), por lo que la utilización de más de una variable resultado plantea la cuestión sobre si, manteniendo ese nivel en cada contraste (0.05), en realidad la probabilidad de encontrar un resultado estadísticamente significativo no es entonces mayor que 0.05. Para explicarlo con términos sencillos: no es igual la probabilidad de que nos toque la lotería si jugamos con un número que si tenemos 10 números diferentes (aunque en este ejemplo probablemente la diferencia real es muy pequeña, salvo en lo que nos cuestan los billetes).

Este problema se conoce en estadística con el nombre de comparaciones múltiples.

Cuando se efectúa más de un contraste estadístico en el análisis de los datos, el criterio aplicado por la mayoría de los investigadores es el de "ajustar" o "corregir" el nivel de corte (inicialmente $p < 0.05$) dependiendo del número de contrastes efectuado. El razonamiento es el siguiente: Si la hipótesis nula (igualdad de los tratamientos) es en realidad correcta (esto no lo sabemos y es lo que intentamos averiguar con nuestro estudio) y usamos un nivel de significación alfa=0.05, para rechazarla estamos, por tanto, declarando una probabilidad de 0.95 de aceptarla siendo cierta (quiere esto decir que si realizásemos muchos estudios del mismo tipo estaremos aceptando −por término medio− la hipótesis nula −siendo cierta− en 95 de cada 100 estudios). Si efectuamos dos pruebas diferentes e independientes para contrastar la hipótesis nula, hay cuatro posibles resultados: ninguna de las dos es estadísticamente significativa (en ambas obtenemos $p > 0.05$), las dos son significativas (ambas con $p < 0.05$) y una de ellas es significativa y la otra no. La probabilidad de que

ninguna de las dos sea significativa, al considerar sucesos independientes, es 0.95x0.95 = 0.90. Luego la probabilidad de que al menos una de ellas sea significativa (suceso complementario) es 1 − 0.9 = 0.1. Por lo tanto la probabilidad global de que en nuestro estudio rechacemos una hipótesis nula es 0.1 y no 0.05 como deseábamos: tenemos dos billetes de lotería.

En el caso de que se efectúen K contrastes la probabilidad de que ninguno sea significativo será $(1-alfa)^K$ y por tanto la probabilidad de que al menos uno de los K contrastes sea significativo es $1-(1-alfa)^K$. Por ejemplo para K=10 contrastes, si utilizamos un alfa=0.05, la probabilidad de que al menos uno de ellos sea significativo es de 0.40.

Este razonamiento conduce al método de ajuste para contrastes múltiples más utilizado rutinariamente y que se conoce como método de Bonferroni, de tal manera que si se efectúan K contrastes, para mantener la probabilidad global ($p < 0.05$) de rechazar incorrectamente en nuestro estudio la hipótesis nula, el nivel de corte a utilizar en cada contraste debe ser 0.05/K. Por ejemplo para 3 contrastes el nivel de P para el que se rechazará la hipótesis nula será de 0.017 (que es bastante más exigente que 0.05). El objetivo de la correción de Bonferroni es por tanto no aumentar la probabilidad global de hallar resultados sólo por el mero hecho de efectuar muchos análisis en diferentes variables obtenidas en nuestra muestra.

Este asunto de las comparaciones múltiples ha sido objeto, y sigue siéndolo, de bastante polémica, y también ha dado lugar en un amplio sector, a la utilización de forma acrítica de la corrección de Bonferroni. En la sección Objetivos múltiples de este artículo profundizaremos algo más en este asunto.

Análisis de subgrupos

El análisis de subgrupos es uno de los casos de multiplicidad que con más frecuencia nos encontramos en la literatura médica, quizás por razones naturales, ya que en principio parece totalmente legítimo y razonable investigar si las diferencias entre los tratamientos dependen de las características de los pacientes pero, como veremos, basarse únicamente en los valores de la probabilidad obtenida para los diferentes subgrupos puede ser cuando menos engañoso. En el caso de que el resultado global sea significativo es casi seguro que encontraremos diferencias significativas y no significativas entre diferentes subgrupos, y con un resultado global no significativo, es probable que encontremos diferencias significativas entre algunos subgrupos por puro azar y esta probabilidad aumentará a medida que aumentemos el número de subgrupos (más billetes de lotería) y por tanto el número de comparaciones. Hay que ser muy cauteloso a la hora de interpretar resultados estadísticamente significativos en alguno de los subgrupos cuando el resultado global no lo fue y muchísimo más si éstos se definen a posteriori de entre un conjunto amplio de comparaciones.

El problema, desde un punto de vista estadístico, en la interpretación de los resultados de un estudio de investigación, radica que a menudo nos olvidamos de que la utilización de pruebas estadísticas formales no nos garantiza que los resultados observados no sean debidos única y exclusivamente a la casualidad y que éstas únicamente nos proporcionan una medida de la probabilidad de que eso haya sido así, probabilidad que nos tranquiliza si es baja, pero probabilidad pequeña no es igual a imposible, y esto es válido en general, no sólo en cuanto al análisis de subgrupos. No nos olvidemos que p=0.05 significa que hay 1 entre 20 posibilidades de estar rechazando la hipótesis nula cuando ésta es en realidad correcta (a la lotería jugamos con probabilidades muchísimo más bajas). Normalmente cuando analizamos los resultados de nuestros trabajos de investigación o cuando leemos los de otros autores, o las revisiones sistemáticas (meta−análisis) no pensamos que quizás los resultados hayan sido obtenidos por pura y simple casualidad, a no ser en aquellas ocasiones en las que estamos predispuestos en contra de la teoría presentada. Está bien buscar explicaciones a resultados sorprendentes o inesperados −así avanza la ciencia−, pero conviene no dejar siempre de lado el azar: medirlo no significa que haya sido eliminado. Cuando gastamos tiempo, esfuerzo y dinero en realizar un estudio de investigación y analizamos los resultados, nos cuesta resignarnos a admitir que no hemos encontrado nada significativo y nos gana la tentación de empezar una "cacería de resultados" entre los posibles subgrupos, definibles en función de los datos basales de los pacientes incluidos en el trabajo.

La dificultad en la interpretación de los resultados obtenidos en ese análisis de subgrupos se agrava por la incorrecta aplicación de pruebas estadísticas y por un ausencia de utilización crítica y meditada de éstas, así como un cuidadoso análisis de los resultados numéricos obtenidos. Un procedimiento (incorrecto) utilizado asiduamente para analizar los datos por subgrupos consiste en calcular el resultado en cada subgrupo y comparar los valores de probabilidad P obtenidos, cuando lo que debiera considerarse es la magnitud del efecto, ya que el valor de la probabilidad lo único que en realidad cuantifica es la precisión en la estimación de ese efecto en el subgrupo y depende tanto del valor medio como de su variabilidad (desviación típica) y del tamaño del subgrupo, que en ocasiones pueden estar muy descompensados, ya que no fueron fijados por diseño.

Vamos a poner un ejemplo con datos ficticios para ilustrar este aserto y recomendamos la lectura de los enlaces sobre interacción citados en las referencias, que son de muy sencilla comprensión y además se ilustran con datos reales de trabajos publicados, tanto para datos numéricos continuos como para proporciones.

En la comparación de dos grupos de tratamiento obtenemos los siguientes resultados

	Tratamiento 1	Tratamiento 2
Media	7.02	6.77
Desviación típica	1.25	1.37
Tamaño	466	774
	Comparación entre tratamientos	
Diferencia	0.25	
Error estándar	0.076	
P	0.001	

Se efectúa un analísis según dos subgrupos, que vamos a llamar A y B, donde los resultados son los que se indican seguidamente

	Subgrupo A		Subgrupo B	
	Tratamiento 1	**Tratamiento 2**	**Tratamiento 1**	**Tratamiento 2**
Media	7.35	7.23	6.9	6.6
Error estándar	0.108	0.096	0.066	0.057
Tamaño	128	204	338	570
Diferencia	0.12		0.3	
Error estándar	0.15		0.087	
P	0.42		0.00057	

Si nos fijamos sólo en los valores de P vemos que hay un efecto global significativo (P=0.001) y que también la diferencia es significativa en el subgrupo B (P=0.00057) pero no lo es en el subgrupo A (P=0.42).

Pero si no nos quedamos sólo en la comprobación de las P, vemos que en el subgrupo B la diferencia es mayor que en el A (0.3 − 0.12 = 0.18) aunque también hay muchos más pacientes en ese subgrupo (908 pacientes en B frente a 332 en A, casi tres veces más), lo que hace que la estimación del efecto en el subgrupo

B sea mucho más precisa y por tanto contribuya a bajar radicalmente el valor de la P en ese subgrupo. Es importante comprobar que, sin embargo, el sentido de la diferencia es el mismo en ambos, y lo que realmente interesa saber es si es significativo el cambio en la magnitud del efecto al pasar desde el subgrupo A (D=0.12) al subgrupo B (D=0.3). Esto se conoce como interacción.

La diferencia entre los subgrupos es 0.18, si se calcula el error estándar de esa diferencia el valor que se obtiene es 0.17, por lo que el valor del cociente 0.18/0.17 = 1.06 nos permite, acudiendo a las tablas de la distribución normal (o t de Student si la muestra fuera más pequeña), conocer la probabilidad de obtener un valor igual o mayor que ése, suponiendo que no hubiera diferencias entre los subgrupos (hipótesis nula). El valor de P que obtenemos −para un contraste bilateral− es 0.3, por lo que NO hay justificación probabilística suficiente para proclamar que exista diferencia entre los subgrupos, como quizás inicialmente nos habíamos precipitado a afirmar.

En uno de los enlaces de interés, que es muy cortito y de agradable lectura, se comentan diferentes ejemplos reales de interpretaciones erróneas de los resultados obtenidos en análisis de subgrupos, que han conducido a decisiones que pueden haber perjudicado a los pacientes.

En ese enlace se presenta también un análisis del estudio ISIS (International Study of Infarct Survival trials) según los subgrupos definidos de acuerdo al signo astrológico de los pacientes (12 subgrupos). En ese estudio el resultado global es altamente significativo ($p < 0.00001$) y sin embargo al efectuar la división en esos 12 subgrupos se encontró dos signos del zodíaco (Géminis y Libra) en los que, por el contrario, el resultado del tratamiento era adverso aunque no significativo. Dice el autor que por supuesto la mayor parte de los médicos (¡aunque no todos!) se rieron cuando se presentaron estos resultados, y sin embargo ¿adoptamos ese sano escepticismo cuando se efectúan presentaciones de otro tipo de resultados más plausibles, pero obtenidos con una metodología similar?. Bien es verdad que esto es utilizar un enfoque bayesiano de la cuestión, lo que siempre hacemos en la vida, ya que en realidad siempre tenemos una hipótesis a priori (y en el caso de la astrología espero que ésta sea de total escepticismo).

Objetivos múltiples

Un problema similar al del análisis de subgrupos se presenta, como ya comentamos al principio del artículo, cuando determinamos en nuestro estudio más de una variable resultado, que en general no serán independientes. Por ejemplo se analiza la presión sistólica y la diastólica.

Sin embargo si meditamos algo más sobre el asunto tampoco las cosas son tan sencillas y conviene que al igual que ha pasado con el valor $p < 0.05$, no sacralicemos la corrección de Bonferroni. Existe una corriente de opinión contraria a la utilización de este tipo de ajuste, fundamentalmente para el caso de múltiples resultados; aunque en principio consideran adecuado efectuar ese ajuste cuando se repite una misma prueba en diferentes subgrupos o en los análisis secuenciales.

El razonamiento que plantean es el siguiente: no parece lógico que la interpretación de los resultados observados sea diferente según el número de pruebas estadísticas que se lleven a cabo. Así el autor del artículo antes referenciado dice que, por ejemplo, en la comparación de dos tratamientos de quimioterapia no es lógico que pueda considerarse como estadísticamente significativa o no la diferencia en las tasas de remisión, dependiendo de si se analiza o no también la tasa de supervivencia, la calidad de vida o las tasas de complicaciones. Dicho de forma intuitiva: si tengo un billete para la lotería primitiva y otro para las quinielas la probabilidad de que me toque la lotería primitiva es la misma que si no hubiera hecho la quiniela: evidentemente lo único que ha variado es la probabilidad de que me haga rico, que sí es ahora mayor (aunque, como saben bien los que juegan todas las semanas a estas cosas, sigue siendo extremadamente baja).

Pongamos otro ejemplo: supongamos que el parámetro analizado es la hemoglobina y que encontramos que resulta significativo con una P = 0.045, si se nos ocurre analizar el hematocrito el resultado aplicando la corrección de Bonferroni ya no sería significativo, aunque observemos que el hematocrito también es significativo (lógicamente) con una P similar por ejemplo 0.047. Esta conclusión parece absurda, ya que realmente el hematocrito nos confirma lo que observamos para la hemoglobina.

Lo que falla en primer lugar es que la corrección de Bonferroni se basa en que los contrastes realizados son independientes, y esto no ocurre cuando estudiamos diferentes parámetros en los mismos pacientes, ya que pueden estar correlacionados (muy correlacionados en el caso del hematocrito y la hemoglobina) y por lo tanto si el resultado es significativo en uno de los parámetros, la probabilidad de que también lo sea en el otro aumenta por la presencia de esa correlación.

Se ha descrito en la literatura otras pruebas de ajuste para los resultados de comparaciones múltiples, más adecuadas y no tan restrictivas como la de Bonferroni, y en concreto en alguna de ellas se tiene en cuenta la presencia de correlación entre los resultados. A nuestro juicio la utilización rutinaria de la corrección de Bonferroni protege excesivamente contra la posibilidad de rechazar de forma errónea alguna de las hipótesis nulas, a costa de disminuir gravemente la potencia de la prueba, sobre todo en aquellos casos en los que se determina un número importante de parámetros, por lo que parece sensato utilizar otros métodos de ajuste más adecuados como los citados en los enlaces y siempre razonar sobre lo que se está haciendo.

Comparación de valores basales

Otro entorno en el que a menudo se presentan comparaciones múltiples se da cuando se pretende comprobar la homogeneidad de los distintos grupos de tratamiento en cuanto a los valores basales de los pacientes. Aquí el problema es diferente, ya que si la asignación de los pacientes a los diferentes grupos se efectuó de forma aleatoria y no se hizo "trampa", sabemos que −por diseño− todos los pacientes pertenecen realmente a la misma población y que, si existieran diferencias, éstas han sido debidas a la casualidad; así que lo más sensato es no efectuar esos contrastes, salvo en cuanto a aquellos factores que realmente pudieran guardar relación con la variable respuesta, y en el caso de que tuviéramos la mala suerte de encontrar una falta de balance analizar los resultados, no sólo de forma directa, sino también ajustados en función de esos covariantes. Por ello aunque en las publicaciones es lógico e incluso obligado proporcionar una tabla descriptiva de los valores basales de los pacientes por grupo de tratamiento, ya que proporciona al lector una idea sobre las características de los pacientes a los que se puede extender las conclusiones del trabajo, no es muy sensato sobrecargar dicha tabla con las comparaciones entre ellos.

Enlaces de interés, de acceso libre

- ***Statistics notes: Multiple significance tests: the Bonferroni method***
  J Martin Bland and Douglas G Altman
  BMJ 1995; 310: 170. [Full text]

- **Debate: Subgroup analyses in clinical trials: fun to look at, but don´t believe them**
  Sleight P.
  Curr Control Trials Cardiovasc Med 2000, 1:25−27

- ***What's wrong with Bonferroni adjustments***
  Thomas V Perneger
  BMJ 1998; 316: 1236−1238. [Full text]

- ***Statistics Notes: Interaction 1: heterogeneity of effects***
  Douglas G Altman and John N S Matthews
  BMJ 1996; 313: 486. [Full text]

Statistics Notes: Interaction 2: compare effect sizes not P values
John N S Matthews and Douglas G Altman
BMJ 1996; 313: 808. [Full text]
Statistics notes: Interaction 3: How to examine heterogeneity
John N S Matthews and Douglas G Altman
BMJ 1996; 313: 862. [Full text]

- ***Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic?***
  Nick Freemantle
  BMJ 2001; 322: 989−991. [Full text]

- ***Multiple test procedures other than Bonferroni's deserve wider use***
  Ralf Bender and Stefan Lange
  BMJ 1999; 318: 600a. [Full text]

- ***Other method for adjustment of multiple testing exists***
  Mikel Aickin
  BMJ 1999; 318: 127. [Full text]

- ***The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis?***
  Carl E Counsell, Mike J Clarke, Jim Slattery, and Peter A G Sandercock
  BMJ 1994; 309: 1677−1681. [Abstract] [Full text]

- ***Multiple significance tests***
  Simon Voss and Steve George
  BMJ 1995; 310: 1073. [Full text]

- Adjusted Bonferroni Comparisons

Otros enlaces de interés

- **Some comments on frequently used multiple endpoint adjustment methods in clinical trials.**
  Sankoh AJ, Huque MF, Dubin N.
  Stat Med 1997; 16: 2529−2542 [Medline]

- **Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods**
  Aickin and H Gensler
  Am J Public Health 1996 86: 726−728. [Abstract]

- **Subgroup analysis and other (mis)uses of baseline data in clinical trials**
  Susan F Assmann, Stuart J Pocock, Laura E Enos, Linda E Kasten
  Lancet 2000; 355: 1064-69

- **No adjustments are needed for multiple comparisons**
  Rothman KJ.
  Epidemiology 1990; 1: 43−46 [Medline]

Indice de artículos

Principio de la página