Bioestadística y Epidemiología

COMPARACIONES MÚLTIPLES

Jorge Dagnino S.[1]

- Las comparaciones entre varios grupos, después de hacer un ANOVA, pueden ser *a priori*, planificadas de antemano, o *a posteriori*.
- Las *a priori* exigen definir antes las comparaciones que se harán. La corrección o método de Bonferroni, se usa para ajustar el error dependiendo del número de comparaciones que se harán.
- Las *a posteriori* incluyen el test de Bonferroni, de Tukey, de Dunnet, de Scheffé y de Student-Newman-Keuls. Su uso depende del diseño experimental y exigen igualdad en el tamaño de los grupos. Cuando los grupos son desiguales, la alternativa es el test de Gabriel.

Cuando hay más de dos grupos para comparar, el ANOVA es sólo el primer paso en el estudio de los resultados. Sólo cuando el resultado de un ANOVA es significativo, sugiriendo que hay diferencias entre los grupos, es lícito proceder a averiguar dónde reside la o las diferencias. Esto se realiza con la aplicación de una de diversas pruebas, también pruebas de hipótesis, que se conocen como procedimientos de comparación múltiple. Estos también siguen un raciocinio similar al usado con las otras pruebas que hemos comentado y su finalidad es ajustar las probabilidades a un nivel conocido de α. La necesidad de este ajuste deriva de que cada vez que se hace una comparación de grupos con datos que no son independientes se está aceptando un error α determinado, generalmente 0,05 en medicina; cada comparación va multiplicando este error. Muchos programas computacionales entregan, junto al resultado del ANOVA, el resultado de la comparación entre grupos usando una batería de estas pruebas adicionales señalando la p calculada con cada una. Si no se conocen los supuestos y limitaciones la tentación es obvia: elegir el test que tiene el ansiado asterisco de la significación estadística. Como la corrección de probabilidades dependerá de la situación, del objetivo y planificación del trabajo, no es lo mismo usar indiscriminadamente una u otra prueba de comparación múltiple. Quien así lo hace arriesga aumentar el error tipo I, declarar como significativa una diferencia cuando en realidad las cifras no permiten hacer tal afirmación, o bien un error de tipo II, declarar no significativa la diferencia cuando en realidad sí lo es, desde un punto de vista estadístico, cuando lo que sucede es que la prueba usada no era la adecuada por no ser la de mayor potencia para la situación en cuestión. Vale la pena reiterar que este tipo de error sólo adquiere importancia cuando la diferencia tiene significación clínica lo que, como hemos dicho antes, debe ser siempre la principal consideración.

Todos los métodos de corrección para comparaciones múltiples tienden a ser conservadores; esto significa que si bien el ANOVA dio un F significativo, el análisis posterior puede no arrojar ningún resultado significativo. Cuando las comparaciones se hacen en un modelo ortogonal, uno de los tipos de diseño experimental, los datos comparados son independientes entre sí por lo que no usan la misma información y no sería necesario ajustar el nivel de α y las comparaciones se podrían hacer directamente. La mayor parte de las veces, sin embargo, los datos son relacionados de una o de otra manera por lo que es indispensable algún tipo de ajuste en el análisis. Tampoco es indiferente qué comparaciones finalmente se hacen. En general, es mejor tener especificado *a priori* qué comparaciones se realizarán pues esto limita el número de ellas y, estrictamente, no sería necesario hacer un ANOVA primero aunque sí ajustar el error α. Otra posibilidad, especialmente cuando no se sabe el tenor de las diferencias y su dirección, es el de tener la libertad de hacer las comparaciones *a posteriori*, aunque ciertamente sin elegir cuáles teniendo a la vista los resultados. La tentación aquí, y que arrasa con

[1] Profesor Titular
División de Anestesiología. Pontificia Universidad Católica de Chile.

todo el raciocinio estadístico, es elegir para comparar aquellas medias que aparecen más lejanas.

COMPARACIONES *A PRIORI* O PLANEADAS

1. Corrección o método de Bonferroni o ajuste del nivel descendente. Al hacer varias comparaciones planeadas se puede compensar el aumento de la probabilidad de obtener significación por azar, disminuyendo el nivel de α requerido para dar por significativo el resultado. Ello se logra dividiendo el valor de α por el número de comparaciones planeado. Por ejemplo, si se planean 5 comparaciones y se desea mantener el nivel de α en el 0,05, se divide este valor por 5 dando 0,01; cada comparación debe ser significativa al nivel de 0,01 para declararla estadísticamente significativa. Es el más conservador de todos.
2. Procedimiento de comparación múltiple de Dunn, también llamado procedimiento t de Bonferroni. Conceptualmente es similar al anterior. Es un método que aumenta el valor crítico de F. La magnitud del aumento depende del número de comparaciones y del tamaño de la muestra.

COMPARACIONES *A POSTERIORI* O NO PLANEADAS

Se hacen después que un ANOVA ha dado un valor de F significativo. Cuál prueba usar dependerá del diseño experimental. Las siguientes se usan para grupos de igual tamaño.

1. Prueba HSD de Tukey (diferencia honestamente significativa). Es aplicable sólo a datos pareados, permitiendo la comparación entre todos los pares de medias. Sería el procedimiento más potente y exacto para usar en estas circunstancias y permite el cálculo de intervalos de confianza. Se dice que el test Tukey es exacto en el ajuste de α a 5% mientras que el Bonferroni es aproximado, en el ajuste de α a 5% o menos.
2. Procedimiento de Scheffé. Es el más versátil de todos los procedimientos usados *a posteriori* pues permite efectuar todo tipo de comparaciones y no sólo las pareadas como el método anterior. Esta flexibilidad implica, sin embargo, un valor crítico más elevado para determinar la significación lo que equivale a una menor potencia. Es, por lo tanto, el más conservador de los métodos de comparación *a posteriori*. Al igual que con la prueba HSD de Tukey, también se pueden calcular intervalos de confianza.
3. Procedimiento de Newman-Keuls o de Student-Newman-Keuls. También sólo puede ser usado para medias pareadas. El resultado es idéntico al de Tukey con dos pasos, es decir, con dos comparaciones entre medias. Con más comparaciones, puede hacer declarar significativa una diferencia que quizás no lo hubiese sido con el Tukey. Vale decir tiene más potencia. Su desventaja es que no permite calcular intervalos de confianza.
4. Procedimiento de Dunnet. Sólo es aplicable en la comparación de varias medias con un grupo control, no permitiendo comparaciones entre las mismas medias. El método es exacto cuando el tamaño de las muestras es igual. En caso contrario, se necesitan otras tablas o se puede aplicar el método de Bonferroni y también el de Newman-Keuls, perdiendo en ambos casos potencia.

Cuando la dimensión de los grupos es desigual las alternativas son menos; una de ellas es el test de Gabriel.

Existen otras pruebas, que figuran con frecuencia en los programas computacionales, que no son recomendables porque no corregirían del todo la distorsión de las comparaciones múltiples y, por ende, diferencias muy pequeñas pueden ser declaradas como significativas. Una es la prueba del rango múltiple de Duncan, que usa el mismo principio que la de Newman-Keuls, y la otra es la prueba de la diferencia significativa mínima (LSD).

REFERENCIAS

1. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991, pp 10-18.
2. Bland M. An Introduction to Medical Statistics. OUP: Oxford. 3rd Ed, 2006.
3. Portney LG, Watkins MP. Foundations of Clinical Research. Applications to practice. 2nd ed. Prentice-Hall: Upper Saddle River, 2000.
4. Bland JM, Altman DG. Statistics Notes: Multiple significance tests: the Bonferroni method. BMJ 1995; 310: 170.
5. Gabriel KR. Simultaneous Test Procedures - Some Theory of Multiple Comparisons. Ann. Math. Statist. 1969; 40: 224-250.

Correspondencia a:
Dr. Jorge Dagnino S.
jdagnino@med.puc.cl