LA UTILIZACIÓN DE LOS PROCEDIMIENTOS DE COMPARACIONES MÚLTIPLES EN LA INVESTIGACIÓN EDUCATIVA EN ESPAÑA

A. Sáez, J. M. Suárez, F. Aliaga y R. M. Bo
Universitat de València*[1]

1. PRESENTACIÓN DEL TRABAJO

Un problema común al que nos podemos enfrentar en cualquier investigación es querer comparar más de 2 grupos de datos para detectar posibles diferencias entre ellos. La utilización de modelos de ANOVA puede permitirnos detectar diferencias, a nivel global, entre las medias involucradas, pero en muchas ocasiones deseamos trabajar a un mayor detalle y detectar las diferencias entre grupos concretos lo que sólo es posible mediante el uso de los Procedimientos de Comparaciones múltiples (PCM).

Las propiedades estadísticas de los PCM son bien conocidas (Miller, 1981; Hochberg y Tamhane, 1987) y el comportamiento de estas pruebas en distintas situaciones ha sido descrita tanto en estudios teóricos (Tukey, 1953; Einot y Gabriel, 1975; Stoline y Ury, 1979; Scheffé, 1970) como en estudios utilizando estrategias de simulación de Monte-Carlo (Dunnett, 1980a; Dunnett, 1980b; Wang, 1971; Maxwell, 1980; Keselman y Rogan, 1978; Keselman, Toothaker y Shooter, 1975; Keselman, Murray y Rogan 1976; Berhardson, 1975; Boardman y Moffitt, 1971). A pesar de todo esto, las Comparaciones Múltiples no suelen ser utilizadas por los investigadores o, en algunas ocasiones se utilizan incorrectamente.

La posibilidad de efectuar comparaciones múltiples ha recibido críticas importantes desde diversas perspectivas (Dawkins, 1983; O'Neill, y Wetherill, 1971; Perry, 1986). Bastantes críticas provienen de la pugna teórica entre las diversas concepciones y corrientes estadísticas. Otras se refieren a principios de utilización incorrecta. Por ejemplo, Wilcox (1987) afirma que solamente cuando se desean realizar todas las comparaciones por pares y si se quiere mantener la probabilidad del error de tipo I igual a a, es perfectamente legítimo omitir la prueba F y usar uno de los procedimientos de contraste de medias. La razón aducida es que la aplicación de las pruebas de comparación múltiple únicamente después de una prueba F significativa es una estrategia que reduce la potencia y el nivel a en una cantidad difícil de determinar.

---

* Avda. Blasco Ibáñez, 21, 46010-VALENCIA. Teléfono 3864430. Centralita 3864420, ext. 6245.

1 Por lo que se refiere a manuales, solamente se cuenta con algunos, el trabajo primigenio de síntesis de Miller, ya actualizado (Miller, 1981), y las recientes exposiciones en los textos de Klockars y Sax (1986) y Toothaker (1993) en la colección de Sage y el más completo de Hochberg y Tamhane (1987); ninguno de ellos traducido al castellano.

La utilización incorrecta de Comparaciones Múltiples lleva a los investigadores a conclusiones erróneas que se reflejan en ambos tipos de error (tipo I y tipo II). En un trabajo de Coward (1991) sobre la utilización de las pruebas de comparaciones múltiples en Estados Unidos se detectan cuatro posibles situaciones que pueden conducir a error en la aplicación de las pruebas: 1) utilizar pruebas de comparaciones de pares cuando lo correcto es utilizar contrastes polinómicos, 2) usar comparaciones múltiples a posteriori en lugar de a priori; 3) utilizar medias aritméticas en lugar de minimo cuadráticas y 4) utilizar una prueba demasiado «liberal».

A estos problemas debidos a la mala utilización, hay que añadir la falta de uso de este tipo de pruebas que, como veremos, se da en nuestro país y que a nuestro juicio se da por dos circustancias:

1) La falta de claridad en los textos de estadistica sobre los distintos procedimientos de comparaciones múltiples con una clara clasificación de las pruebas en sus aspectos más relevantes como: conveniencia respecto al diseño experimental utilizado, tratamiento del control del error de tipo I, o incluso a nivel de los supuestos estadisticos necesarios para su aplicación . Y como consecuencia de lo anterior,

2) La falta de implementación en paquetes estadísticos de ordenador de muchas de las pruebas para los diseños experimentales más utilizados. Aunque todos los paquetes suelen cubrir los diseños de una via entre grupos, hemos detectado una carencia de pruebas de comparaciones múltiples para diseños factoriales de medidas repetidas o en los diseños mixtos o en los diseños de ANOVA no paramétricos.

Existen diferentes tipos de clasificaciones de las pruebas de comparaciones múltiples lo cual nos ofrece una variedad de dimensiones según las cuales caracterizar a las pruebas (Hochberg y Tamhane, 1987; Toothaker, 1991). Por ejemplo, Toothaker (1991, 1993) propone las siguientes dimensiones: 1) **Según el número de comparaciones**; 2) **según el tipo de contraste** (ortogonales frente a no ortogonales); 3) **según la manera de comparación** (por pares o no); 4) **comparaciones a priori o comparaciones a posteriori**; 5) **según el proceso de cálculo** (simples o en un único paso o en varios pasos 'stepwise', estos a su vez se dividen en *step-down* o *step-up*, según se proceda desde la mayor diferencia hasta la más pequeña o desde la menor diferencia a la mayor); 6) **según el tipo de estadístico y/o la distibución teórica utilizada en su cálculo**; 7) **según el tipo de tasa de error** (existen dos tipos: tasa de error por comparación y tasa de error por familia).

Muchas de las dimensiones anteriores pueden aparecer combinadas según la elección de la prueba que se realice. Se podrían utilizar comparaciones múltiples a priori y una tasa de error por comparación. O se pueden utilizar comparaciones ortogonales y a posteriori con una tasa de error por comparación. Algunas combinaciones son imposibles, tales como todas las comparaciones por pares y ortogonales. Sin embargo, es cierto que unas determinadas combinaciones se suelen utilizar con más frecuencia, como comparaciones ortogonales y a priori con una tasa de error por comparación.

A modo de cuadro-resumen presentamos algunas de las pruebas de comparaciones múltiples clasificadas según la distribución estadística que utilizan en su cálculo:

Basadas en la distribución t	Dunn-Bonferroni (Dunn, 1961) Dunn-Sidak (Dunn, 1958 y Sidák, 1967) Holm-Shaffer (Holm, 1979 y Shaffer, 1986)
Basadas en la distribución del Rango Studentizado	Tukey (Tukey, 1953) Newman-Keuls (Newman, 1932 y Keuls 1952) Duncan (Duncan, 1955) Ryan (Ryan, 1960; Einot y Gabriel , 1975) Peritz (Peritz, 1970)
Basadas en la distribución F	Scheffé (Scheffé, 1953, 1959) F de Newman-Keuls F de Ryan
Basadas en una prueba t protegida	LSD de Fisher (Fisher, 1935) Shaffer-Ryan (Shaffer, 1979) Fisher-Hayter (Hayter, 1986)
Basadas en la comparación con un control	Dunnet (Dunnett, 1955)

En el resumen anterior podemos observar que, además de las pruebas tradicionales, existen otros procedimientos más recientes (como los de Ryan o Peritz) que suelen ser modificaciones de pruebas anteriores para corregir algunos de los problemas de estas pruebas clásicas. Se da la circustancia que muchas de estas pruebas recientes no están disponibles en los manuales de estadistica al uso y, por supuesto, tampoco estan implementadas en los paquetes estadísticos para ordenador más utilizados.

Como ocurre en la prueba t y la prueba F es necesario el cumplimiento de los supuestos paramétricos para la correcta aplicación de las pruebas de contraste que hemos visto. Vamos a revisar aquí algunos de los resultados más importantes que se han encontrado sobre el cumplimiento de supuestos.

Respecto al supuesto de normalidad parece que según Dunnet (1982), y como ocurre con la prueba F, las pruebas de contraste son robustas frente a pequeñas desviaciones respecto a la normalidad; en el caso de desviaciones de la normalidad muy grandes, aumenta el riesgo del error (consultar a Dunnet, 1982 y Ringland, 1983).

Respecto al problema de tamaños muestrales distintos entre los grupos a comparar se proponen distintas pruebas alternativas como son dos variaciones a la prueba de Tukey: una propuesta por Kramer (1956), conocida como prueba de Tukey-Kramer, y la segunda variación propuesta por Miller (1981) y Winer (1971), conocida como prueba de Miller-Winer y por último una prueba nueva propuesta por Hochberg (1974) conocida como la prueba GT2.

Mientras que para el caso de desigualdad de varianzas existen alternativas como la prueba GH de Games y Howell (1976), y las pruebas C o T3 de Dunnet (1980). Brown y Forsythe (1974c) han propuesto una modificación a la prueba de Scheffé para hacerla resistente a la desigualdad de las varianzas y que ha dado pie a posteriores variaciones (Kaiser y Bowden, 1983) y alternativas (Dalal, 1975; Hochberg, 1976).

En este trabajo se pretende llevar a cabo un contraste entre estas posibilidades técnicas y la realidad de utilización concreta en nuestro ámbito de investigación. Así, pretendemos determinar en que situaciones se emplean estos procedimientos, en cuales se podrían/deberían emplear, qué opciones

concretas se manejan, etc. Todo ello vamos a llevarlo a cabo dentro del ámbito de la investigación educativa en nuestro país.

2. MÉTODO

A tal efecto se ha tomado como referente de investigación los trabajos publicados en revistas de investigación de difusión nacional. En este sentido, se han seleccionado 7 publicaciones: Revista de Investigación Educativa, Infancia y Aprendizaje, Revista de Educación, Investigación en la Escuela, Bordón, Revista Española de Pedagogía y Ciencias de la Educación. Dada la relativa recencia de estos procedimientos como tema monográfico en la literatura estadística se ha seleccionado un conjunto de 5 años correspondientes al período que va desde 1988 a 1992, ambos inclusive.

La técnica de trabajo es la correspondiente a cualquier estudio bibliométrico básico, procurando mantener en todo momento la conexión con las informaciones cualitativas que se derivan de los informes originales.

3. RESULTADOS

El conjunto de resultados respecto a la utilización de las pruebas de Comparaciones Múltiples se encuentran recogidos en las tablas 1 y 2. Es preciso resaltar que se indagan únicamente tres publicaciones (Revista de Investigación Educativa, Infancia y Aprendizaje y Bordón) por ser las únicas de las revistas estudiadas en las que se encuentran artículos que utilizan las pruebas de Comparaciones Múltiples.

A partir del análisis de esta información se pueden señalar los siguientes aspectos relevantes:

- Existe esencialmente una revista —Infancia y Aprendizaje— en la que es relativamente habitual la publicación de trabajos que incluyen la utilización de estas pruebas. Y aún así su presencia es relativamente moderada respecto a las posibilidades potenciales directas para su empleo: un 38,89% de los artículos que utilizan modelos ANOVA. En las otras dos revistas la publicación de trabajos que presenten Comparaciones Múltiples es mucho más rara.
- No parece existir ningún tipo de evolución temporal a través de los cinco años estudiados respecto a la utlización de las pruebas de Comparación Múltiple. De hecho, los escasos trabajos que las utilizan tienden a distribuirse de una forma casi uniforme a través de todo este período temporal.
- Merece un comentario detallado la escasa utilización de estas pruebas incluso tomando como referente las situaciones en las que es perfectamente ajustada su utilización: aquellos trabajos en los que se ha utilizado modelos de ANOVA. Así, sólo un 13,64% de los trabajos de las tres revistas, entre los que emplean modelos ANOVA, ha utilizado alguna técnica de Comparaciones Múltiples. Además, sí que parece haber una cierta relación entre la mayor utilización de estas técnicas y su mayor actuación proporcional respecto al total de situaciones posibles. Así, en ambos casos la revista Infancia y Aprendizaje muestra tanto una mayor frecuencia de utilización como una mayor proporción de utilización respecto a todos los trabajos que emplean modelos ANOVA.

TABLA 1

NÚMERO DE ARTÍCULOS CON APLICACIONES DE PRUEBAS DE COMPARACIONES MÚLTIPLES PARA CADA REVISTA Y AÑO. SE ESTABLECEN PROPORCIONES RESPECTO A LA TOTALIDAD DE LOS ARTÍCULOS PUBLICADOS, RESPECTO A LA TOTALIDAD DE ARTÍCULOS DE LA REVISTA Y RESPECTO A LA TOTALIDAD DE ARTÍCULOS POR AÑOS —EN LOS TRES CASOS ES LA TOTALIDAD DE ARTÍCULOS QUE UTILIZAN ESTAS PRUEBAS—. (T= TUKEY, S=SCHEFFÉ Y B= BONFERRONI)

1988 1989	1990	1991	1992	Total		
Revista Investigación Educativa	0	0	0	1(T)	0	1
% respecto Total	0,0	0,0	0,0	11,11%	0,0	11,11
% respecto Revista	0,0	0,0	0,0	100%	0,0	
% respecto Año	0,0	0,0	0,0	33,33%	0,0	
Infancia y Aprendizaje	2(T)	1(S)	1(B)	2(S)	1(S)	7
% respecto Total	22,22	11,11	11,11	22,22	11,11	77,78
% respecto Revista	28,57	14,29	14,29	28,57	14,29	
% respecto Año	100	100	50	66,67	100	
Bordón	0	0	1(S)	0	0	1
% respecto Total	0,0	0,0	11,11	0,0	0,0	11,11
% respecto Revista	0,0	0,0	100	0,0	0,0	
% respecto Año	0,0	0,0	50	0,0	0,0	
Total	2	1	2	3	1	9
% Total	22,22	11,11	22,22	33,33	11,11	

- Los trabajos que emplean Comparaciones Múltiples tienden a utilizar aquellas que se corresponden con los modelos más simples —ANOVAS de una vía— con técnicas recogidas en los paquetes estadísticos más extendidos (SPSS, BMDP, fundamentalmente). Una parte de las «exclusiones» en cuanto a la utilización de estos procedimientos se produce por la presencia de modelos ANOVA más complejos —con 2 ó más variables independientes—. Esto parece tener una clara relación con la mucho menor oferta de estos procedimientos en los paquetes estadísticos, junto con un tratamiento más esporádico del problemas en los textos y manuales dedicados a esta temática. Así, salvo los manuales que tratan monográficamente el tema de las técnicas de Comparaciones Múltiples, las opciones y forma de aplicación de las mismas a los modelos de cierta complejidad no están tradicionalmente recogidas en los textos, salvo alguna mención tangencial en algunos casos.

TABLA 2

NÚMERO DE ARTÍCULOS CON APLICACIONES DE PRUEBAS DE COMPARACIONES MÚLTIPLES PARA CADA REVISTA, PROPORCIONES SOBRE EL TOTAL DE ARTÍCULOS QUE HAN EMPLEADO UN MODELO ANOVA Y SOBRE EL TOTAL DE ESTUDIOS EN QUE SE UTILIZA EL ANÁLISIS CUANTITATIVO

	1988	1989	1990	1991	1992	Total
Revista Investigación Educativa	0	0	0	1(T)	0	1
% respecto Trabajos estadísticos	0,0	0,0	0,0	11,11%	0,0	1,33
% respecto ANOVAS	0,0	0,0	0,0	100	0,0	8,33
Infancia y Aprendizaje	2(T)	1(S)	1(B)	2(S)	1(S)	7
% respecto Trabajos estadísticos	14,29	5,88	8,33	14,29	11,11	10,61
% respecto ANOVAS	100	12,5	25	66,67	100	38,89
Bordón	0	0	1(S)	0	0	1
% respecto Trabajos estadísticos	0,0	0,0	16,67	0,0	0,0	1,72
% respecto ANOVAS	0,0	0,0	100	0,0	0,0	9,09
Total	2	1	2	3	1	9
% respecto Trabajos estadísticos	3,45	1,12	1,9	5	2	2,49
% respecto ANOVAS	22,22	6,67	8,7	30	11,11	13,64

- Del total de 9 artículos en los que se emplean técnicas de Comparaciones Múltiples más de la mitad (55,56%) utilizan la prueba de Scheffé. En los casos restantes, se utilizan en una tercera parte (3 artículos) la prueba de Tukey (%) y en un sólo artículo (%) la prueba de Bonferroni. A partir de las informaciones sobre los estudios respecto a las propiedades de estas pruebas las decisiones no parecen estar mayoritariamente respaldadas por la evidencia. Toothaker (1993) recomienda la utilización de la prueba de Tukey por no resultar tan conservadora como la de Scheffé ni tan liberal como la de Bonferroni. Por su parte, HOCHBERG Y TAMHANE (1987) señalan que la prueba más potente con diseños equilibrados es la de Scheffé, mientras que la de Tukey es más adecuada cuanto más se acentúa el desequilibrio entre los grupos. Dado que la práctica totalidad de los estudios manejan grupos desequilibrados parece que se ha producido mayoritariamente una decisión no suficientemente avalada, respecto a la técnica concreta elegida para establecer las Comparaciones Múltiples, en dos terceras partes de los trabajos estudiados (66,66%).
- Por lo que se refiere a la utilización de las pruebas dentro del enfoque del ajuste de modelos, es preciso resaltar que se produce en la práctica totalidad de las situaciones una aplicación independiente de la verificación del cumplimiento de los supuestos del modelo. Este hecho, por otra parte, es algo desgraciadamente habitual en el caso de la aplicación de múltiples técnicas estadísticas y particularmente se tiende a producir en los casos en que se utilizan modelos ANOVA.

4. CONCLUSIONES

La revisión de la literatura científica más reciente respecto a las técnicas estadísticas está acentuando cada vez más la importancia de la utilización de los procedimientos de Comparaciones Múltiples cuando se trata de verificar hipótesis sobre la igualdad de K medias correspondientes a situaciones diferentes que se desean comparar.

Existe hoy en día suficientes alternativas desarrolladas dentro de las técnicas de Comparaciones Múltiples que pueden dar respuesta a buena parte de las necesidades más comunes en la investigación educativa. De hecho, se cubre la mayor parte de los modelos que se aplican a las situaciones que habitualmente se plantean en nuestro ámbito de investigación.

No obstante, a partir de la revisión de los trabajos de investigación educativa en el ámbito español se aprecia que, en conjunto, el tema de la aplicabilidad de estas técnicas está poco difundido en la comunidad científica. De hecho, se manejan los procedimientos más clásicos que son de amplia difusión en los textos de estadística y que se encuentran profusamente recogidos en los paquetes estadísticos más importantes. Además, se emplean exclusivamente aquellas técnicas relacionadas con los modelos más simples, con una sola variable independiente y en situaciones entre grupos. Por último, cabe señalar que las decisiones entre las opciones alternativas no se suelen hacer tomando como base los estudios de validación disponibles respecto a los diferentes procedimientos.

Por todo ello, es preciso afirmar la necesidad de establecer los mecanismos para la difusión de este procedimiento entre la comunidad científica que investiga en educación. Esto entendemos que debe hacerse realizando un esfuerzo por clarificar las opciones disponibles, establecer estrategias de adecuación de estas opciones a cada situación concreta y reseñar cuales son las herramientas informáticas disponibles en cada situación.

Finalmente, pensamos que sería de particular interés en este tema la elaboración de programas informáticos específicos que cubrieran las principales lagunas existentes y que permitieran una utilización más amplia y mejor dirigida de estos procedimientos, a la espera de su implementación en los paquetes estadísticos más conocidos.

5. BIBLIOGRAFÍA

BERHARDSON, C. (1975): Type I Error Rates when Multiples Comparison Procedures Follow a Significant F Test of ANOVA. *Biometrics*, **31**, 229-232.

BOARDMAN, T. & MOFFITT, D. (1971): Graphical Monte Carlo Type I Error Rates for Multiple Caomparison Procedures. *Biometrics*, **27**, 728-744.

BROWN, M. & FORSYTHE, A. (1974c): The ANOVA and Multiple Comparations for Data with Heterogeneous Variances. *Biometrics*, **30**, 179-184.

COWARD, W. M. (1991): *A Meta-Analysis of Multiple Comparison Procedures*. Tesis doctoral.

DALAL, S. (1975): Simultaneous Confidence Procedure for Univariate Behrens-Fisher Type Problems. *Biometrics*, **65**, 221-225.

DAWKINS, H. (1983): Multiple Comparisons Misused: Why so Frequently in Response-Curve Studies? *Biometrics*, **39**, 789-790.

DUNCAN, D. (1955): Multiple Rage Tests and Multiple F Test. *Biometrics*, **11**(1), 1-42.

DUNN, O. J. (1958): Estimation of the Means of dependent variables. *Annals of Mathematical Statistics*, **29**, 1.095-1.111.

DUNN, O. J. (1961): Multiple comparisons using rank sums. *Technometrics*, **6**, 241-252.

DUNNETT, C. (1955): A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association*, **50**, 1.096-1.121.