Agronomía Colombiana. 1989. Volumen VI: págs. 19-24

¿Estamos Usando los Procedimientos Adecuados para Comparar Promedios entre Tratamientos?[1]

RICARDO MARTINEZ B.[2]

Resumen. Cuando los experimentadores desean comparar los verdaderos efectos de sus tratamientos, comúnmente recurren a un análisis de varianza, a la correspondiente prueba de F, y por último a alguna de las pruebas de comparación múltiple, para poder hacer recomendaciones específicas. En este proceso no dejan de incurrir en ciertos errores que en algunas oportunidades los llevan a conclusiones equivocadas.

En algunos casos se abusa de las posibilidades que ofrecen las diferentes pruebas, y en muchas otras oportunidades simplemente no se cumplen los supuestos mínimos para su aplicación. Este problema ha sido discutido por algunos autores, quienes no se han puesto de acuerdo en varios aspectos. En este criterio se examinarán críticamente algunas situaciones del uso inadecuado de tales pruebas en el medio colombiano con el propósito de colaborar con los investigadores en el sector agrícola para mejorar su uso en el futuro.

USE AND MISUSE OF MULTIPLE COMPARISON PROCEDURES

Abstract. **When researches want to compare the real effect of their treatments they usually apply analysis of variance, the F test and finally multiple comparison procedures in order to make specific recomendations. In this process they sometimes make mistakes that affect final conclusions. Uses and abuses of multiple comparation procedures are discussed in this paper. If the treatment in an experiment are structured, such as in a factorial arrangement or a quantitative series of levels, the structure should be examined in the analysis. Multiple comparisons tests such as Duncan's Multiple Range Test or a test using Least Significant Difference, assume that the treatments being compared are not structured. If they are used on structured experiments, useful information will be ignored and untenable or biased conclusions mybe reached.**

INTRODUCCION

Colectar datos dentro de las investigaciones cuesta dinero, por lo tanto los investigadores tienen la responsabilidad de gastar el tiempo que sea necesario en el análisis de sus datos. Desafortunadamente un número apreciable de investigadores usan la metodología estadística como un mal necesario, haciendo el menor esfuerzo por sacarle el mayor provecho. Esto explica la marcada tendencia al uso de un número reducido de técnicas estándares que aparecen en los textos o que han usado sus jefes, en vez de ajustarse a los objetivos propuestos en la investigación.

Sabemos bien que el uso inadecuado de la metodología estadística, aún en casos relativamente simples, lleva a interpretaciones inadecuadas de los resultados. Muchos artículos escritos por un número grande de autores estadísticos y no estadísticos, recalcan sobre el uso y el abuso de la bioestadística en la investigación biológica. Además, con frecuencia se presenta el problema de la copia entre diferentes investigadores de análisis aparentemente adecuados, perpetuándose fácilmente técnicas estadísticas inapro-

[1] Recibido para publicación el 14 de agosto de 1989.

[2] Profesor Asociado, Universidad Nacional de Colombia, Facultad de Agronomía, Bogotá, Colombia. A.A. 14490, Bogotá.

piadas. Un ejemplo clásico de esta situación es el mal uso de las pruebas de comparación múltiple entre medias de tratamientos, especialmente cuando se usa la prueba **Rango Múltiple de Duncan** con datos que presentan una estructura definida, tal como ocurre con los experimentos factoriales o aquellos en que los tratamientos forman series progresivas.

Cuando se desea usar una cierta prueba para comparar medias de tratamientos, realmente no importa qué diseño se usó. Sin embargo, esto no quiere decir que la prueba usada corrija los errores cometidos en un experimento inadecuadamente diseñado, mal ejecutado y analizado. Si el control local no se ha hecho apropiadamente, si el número de repeticiones es deficiente y la aleatorización es defectuosa, es natural que todo el proceso se verá afectado negativamente, caso en el cual será imposible que una prueba detecte la real situación de los efectos de los tratamientos. Si se ha escogido y aplicado la metodología adecuada, tanto para el análisis de varianza como para las pruebas de comparación de medias de los tratamientos, debe encontrarse que hay diferencias significativas estadísticamente hablando cuando en las respectivas poblaciones así ocurre, a menos que el nivel del Error Tipo II haya sido realmente grande.

EL PROBLEMA: ¿QUE ES MAS IMPORTANTE, LAS PRUEBAS DE HIPOTESIS O LA ESTIMACION?

Algunos autores consideran que el interés primordial en la agricultura y de la biología en general es la de estimar la magnitud de los efectos de los tratamientos (Yates, 1951, 1964; Cox 1977; *et al* 1978 y Chew 1980), aunque otros como Fisher (1960) y Anscombe (1965) le dan mayor importancia a las pruebas de hipótesis.

Muy seguramente, un investigador en fertilización puede estar más interesado en saber el grado en que aumenta la producción de un cultivo cada vez que aumenta cierta cantidad de un fertilizante, o un fitomejorador puede estar interesado en saber en qué proporción unos genotipos nuevos son mejores que un testigo, en vez de solamente saber si hay o no hay diferencias significativas entre los niveles de los fertilizantes o entre los genotipos. Sin embargo, a pesar de las debilidades que algunos agrónomos y biólogos les achacan a las pruebas de hipótesis no sobra recalcar que éstas juegan un papel muy importante en el método científico. Claro está que hay que diferenciar entre pruebas de hipótesis experimentales y pruebas de hipótesis estadísticas, aunque éstas se usan como herramientas para probar las primeras. Por otro lado, no es conveniente sobre enfatizar el uso de las pruebas de hipótesis estadísticas a expensas de subestimar las hipótesis experimentales, este es un caso frecuente cuando se usan las pruebas de comparación múltiple.

CRITICAS MAS FRECUENTES A LAS PRUEBAS DE COMPARACION MULTIPLE

Como se sabe, cada uno de los métodos de comparación múltiple fue desarrollado para una situación muy particular, por lo general limitada, pero en la práctica se usan muy profusamente sin ninguna consideración previa, respecto a lo apropiado de su aplicación.

Veamos algunas opiniones de destacados estadísticos reunidos con ocasión de la presentación del escrito O'Neill y Wetherill (1971). Nelder, en esa época director del Departamento de Estadística en Rothamsted, considera que las pruebas de comparación múltiple no tienen nada que hacer en la interpretación de datos; va hasta el punto de considerar que sólo sirven para, de manera artificial, darle interés a datos que no lo son. Respecto a la prueba de Duncan varios autores, entre ellos Scheffé, la rechazan de plano porque no entienden su justificación, es más, consideran que hay errores en su publicación original. Los mismos autores O'Neill Wetherill opinan que muchos métodos de comparación múltiple constituyen un intento por amoldarse al tipo de estadística de Neyman-Pearson, la cual en realidad es más flexible y se ajusta más a la forma de ver los datos por parte de los investigadores. Sin embargo, el profesor R.L. Plackett, de

la Universidad de Newcastle, piensa que el experimentador debería usar metodologías más sencillas y menos confusas. Por su parte el profesor D.R. Cox, del Imperial College de Inglaterra, dice que las pruebas de Tukey y Scheffé son valiosas pero conservadoras porque quizá es raro que el investigador encuentre significancia en semejante conjunto de comparaciones que se contemplan en esas pruebas, y cuando se consideran los límites de confianza para las restricciones escogidas. Otros autores sugieren el análisis de conglomerados o metodología bayesiana. Afortunadamente se dispone de la prueba de la Diferencia Significativa Mínima Bayesiana que ofrece una alternativa cuando el experimentador desea hacer una estimación subjetiva de la importancia relativa.

¿QUE HACER ENTONCES?

Las pruebas de comparación múltiple se deben usar con mucha cautela, debe identificarse y estudiarse concienzudamente cada situación. Una prueba específica para comparar medias de tratamientos depende del propósito del experimento y si tales tratamientos tienen estructura, tal como ocurre con los experimentos factoriales y como factores cuantitativos con niveles seriados, tal estructura debe estudiarse en el análisis. Pruebas de comparación múltiple tales como la prueba de Rango Múltiple de Duncan o la de la Diferencia Significativa Mínima, suponen que los tratamientos que se comparan no tienen estructura (Chew, 1976). Si tales pruebas se usan en experimentos con tratamientos estructurados, información útil se ignorará, y puede llegarse a conclusiones sesgadas o insostenibles. Por lo tanto, siempre que se tenga tratamientos con una estructura específica es aconsejable usar **contrastes ortogonales.** Estas se usan para probar y estimar comparaciones específicas, hechas de antemano porque constituyen una técnica poderosa y flexible. La ortogonabilidad maximiza el uso de la información, sin embargo, el experimentador no debe convertirse en un esclavo de la ortogonabilidad hasta el punto de que llegue a probar hipótesis no importantes, desde el punto de vista del área de interés.

En algunas ocasiones los datos pueden sugerir las comparaciones inesperadas resultantes de diferencias grandes entre ciertos promedios, que pudieron no considerarse de antemano. En estos casos pueden usarse pruebas como las de Scheffé, prueba que es más flexible y puede usarse para hacer comparaciones a posteriori entre medias con diferente número de repeticiones. También puede calcularse intervalos de confianza de la estimación de las diferencias entre medias o entre grupo de medias.

ALGUNOS EJEMPLOS

Ejemplo 1. Interpretación de la aplicación de dosis de un producto y su respuesta. Se llevó a cabo un experimento en que se quería probar siete tratamintos que consistían en siete niveles de un factor (Ejemplo: un insecticida, un fungicida, un fertilizante, etc.), aplicado a un determinado cultivo. Una pregunta simple, directa y útil pudo plantearse el experimentador y es la siguiente: ¿Qué relación existe entre el material aplicado y la respuesta de la planta? Además pudo plantearse otra pregunta menos simple y menos útil como ésta: ¿De los posibles 21 pares de tratamientos, cuáles son estadísticamente diferentes? Si esta segunda pregunta fue la que realmente se planteó el experimentador, el siguiente cuadro de resultados sería el adecuado.

Nivel del Factor	Respuesta del Cultivo
0	222 a
1	202 a b
2	205 a b
3	186 b c
4	164 c d
5	156 c d
6	147 d

Ahora supóngase que la pregunta hecha y respondida fue la de la relación entre dosis y respuesta. Entonces se habría reportado que se presentó una relación lineal altamente significativa, en que el 96% de la variación de la respuesta se debió al aumento de las dosis. Esto puede ilustrarse en el Cuadro 1 y Figura 1.

Ahora, la pregunta de cuáles respuestas a los tratamientos son significativamente diferentes es irrelevante, y no se necesita, es más, no debió hacerse. Una vez se establece la significancia de la tendencia lineal dentro del rango usado, eso implica que hay diferencias significativas entre todos los niveles de tratamientos usados. Así, los mejores estimativos de los efectos de los tratamientos son los puntos sobre la línea de regresión.

Naturalmente, existen otros tipos de relación entre dosis y las respuestas además de la lineal, tales como varios tipos de relaciones curvilíneas.

Cuadro 1. Participación de las Sumas de Cuadrados del Ejemplo 1.

Fuentes de Variación	Grados de Libertad	Sumas de Cuadrados	Porcentaje de influencia
Tratamientos	6	23805	
Lineal	1	22885	96%
Residuo	5	920	4%

Ejemplo 2. Comparación entre medias de genotipos que presentan una estructura. Se tienen cinco genotipos de un determinado cultivo y se quiere conocer su adaptabilidad a un nuevo ambiente. Se usa un genotipo (E) estándar como testigo, dos híbridos de crecimiento no indeterminado (Ay B) y dos de crecimiento indeterminado (C y D). Además, los híbridos son de diferente origen unos holandeses, otro japonés y una variedad colombiana (E). Las preguntas que se pueden hacer son las siguientes: ¿Qué tan buenos son los nuevos híbridos (A,B,C y D) comparados con el genotipo nacional (E)? Qué diferencias hay entre los híbridos no indeterminados comparados con los determinados? Qué diferencias hay dentro de los no indeterminados y dentro de los determinados? Por otro lado, simplemente pudo haberse preguntado si hay diferencias significativas entre los 10 pares de tratamientos originados, esto para una determinada variable de respuesta. Si esta última pregunta fue la que se planteó, el siguiente Cuadro de resultados sería la adecuada.

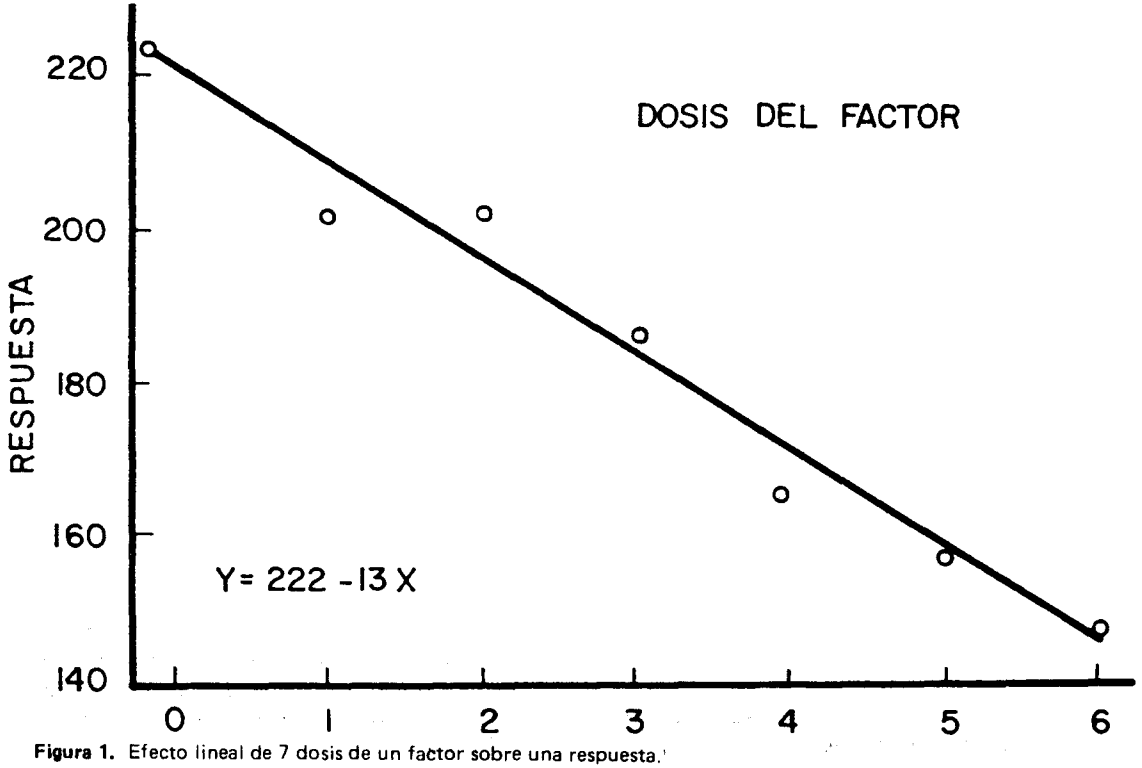

Figura 1. Efecto lineal de 7 dosis de un factor sobre una respuesta.

Tratamiento	Promedio
B	86,512 a
A	75,866 b
C	70,455 bc
D	65,488 c
E	26,693 d

Sin embargo, si la pregunta fue la primera, la presentación debería ser la siguiente (Cuadro 2).

Cuadro 2. Descomposición de Sumas de Cuadrados del Ejemplo 2. Variable 2.

Fuentes de Variación	Grados de Libertad	Sumas de Cuadrados
Bloques	2	
Tratamientos	4	6.234,806 **
E Vs (A,B,C y D)	1	5.503,710 **
(A y B) Vs (C y D)	1	524,173 **
C Vs D	1	37,041 ns
A Vs B	1	170,027 *
Error	8	290,783

** $P < 0,01$
* $P < 0,05$

Acá, en el Cuadro 2, se está examinando la estructura de los tratamientos, se está no solamente comparando los promedios sino discriminando la bondad de los grupos formados en los tratamientos debido a la característica no indeterminado, trabajando con **Contrastes Ortogonales.**

Ejemplo 3. Factoriales. Si se tiene un factorial 2x2x2, o sea, mi experimento donde se tienen 3 factores cada uno con dos niveles. Las preguntas que se han debido hacer son: ¿Cuál es el efecto de cada factor? ¿Existen interacciones entre los factores?

Otra pregunta pudo ser, si existen diferencias significativas entre los posibles 28 pares de comparaciones entre tratamientos, para lo cual la siguiente presentación de resultados es la adecuada.

Como ya se sabe, tratamientos bajo la misma letra no presentan diferencias significativas. Estas letras no nos dicen mucho, pero si partimos las sumas de cuadrados en efectos principales e interacciones, encon-

Combinación de Factores	Respuesta
a_1 b c_1	10,4 b
a_1 b c	8,5 b
a_1 b c_1	15,2 a b
a_1 b c	13,0 a b
a_o b_1 c_1	21,4 a
a_o b_1 c_o	18,6 a b
a_o b_1 c_1	22,6 a
a_o b_o c_o	22,7 a

tramos que el Factor A presentaba un efecto altamente significativo, explicando en un 86% el efecto total de los tratamientos. El factor B explicó el 12% de la variación y no hubo evidencia del efecto del factor C ni de ninguna interacción, tal como se puede apreciar en el Cuadro 3.

Cuadro 3. Repartición de las Sumas de Cuadrados de los Tratamientos. Ejemplo 3.

Fuente de Variación	Grados de Libertad	Sumas de Cuadrados	Porcentaje
Tratamientos	7	877	
Factor A	1	730	83%
Factor B	1	107	12%
Factor C	1	23	3%
AB	1	8	
AC	1	1	
BC	1	3	2%
ABC	1	5	

En cada uno de los ejemplos dados se ha mencionado la repartición o la descomposición de los efectos de los tratamientos como si la técnica para hacerlo fuera de conocimiento común. Desafortunadamente no es así, pues lo común es usar pruebas como la de Duncan o la de la Diferencia Significativa Mínima. Esto no es muy afortunado ya que la técnica que se ha usado en los ejemplos como la apropiada es muy poderosa y es relativamente muy simple. A esta prueba se le conoce como "contrastes ortogonales", o "formas lineales ortogonales" o "grados de libertad individuales".

RECOMENDACIONES

1. Al usar las pruebas de comparación múltiple:
Hágalo luego de un examente cuidadoso y en las situaciones apropiadas.
No lo haga de manera rutinaria, sólo porque un paquete de computación se lo facilita. No lo haga para tratamientos que presenten una estructura.
2. Al planear un experimento, decida de manera clara y definitiva el tipo de preguntas que desea responder y diseñe el experimento que le permita hacerlo.
3. Al presentar los resultados, comuníquele al lector en qué tipo de preguntas se basó el experimento.
4. Interprete los resultados a manera de respuestas a las preguntas que usted se hizo al principio.

LITERATURA CONSULTADA

1. Anscombe, F.B. 1965. Comments on Kurtz-Link-Tukey-Wallace paper. Technometries 7: 167-168.
2. Box, G.E.P., W.G. Hunter and J.S. Hunter 1978. Statistics for experimenters. Wiley, New York.
3. Chew, V. 1976a. Uses and abuses of Duncan's multiple range test. Proc. Fla. State Hortic. Soc. 89-251253.
4. Chew, V. 1980. Testing differences among means: Correct interpretation and some alternatives. Hort Science 15: 467-470.
5. Cox, D.R. 1977. The role of significance tests Scand J. Stat. Theory Appt. 4: 49-70.
6. Miller, R.G. 1978. Simultaneous statistical inference, 2nd. ed. McGraw-Hill Book Co., New York.
7. O'Neill, R. and G.B. Wetherrill. 1971. The present state of multiple comparison methods. J.R. Stat. Soc. (B) 36: 218-250.
8. Scheffé, H. 1959. The analysis of variance. John Wiley and Sons, Inc. New York.
9. Yates, F. 1951. The influence of Statistical Methods for Research Workers on the development of the science of statisties J. Am. Stat. Assoc. 46: 19-34.
10. Yates, F. 1964. Sir Ronald Fisher and the design of experiments. Biometrics 20: 307-321.