

Aplicación de métodos de comparaciones múltiples en Biotecnología Vegetal

Gladys Casas Cardoso^{1*}, Novisel Veitía² * Autor para correspondencia

¹Laboratorio de Bioinformática, Facultad de Matemática, Física y Computación, Universidad Central Marta Abreu de Las Villas. Carretera a Camajuaní km 5.5. Santa Clara. Villa Clara. Cuba. CP 54 830 e-mail: gladita@uclv.edu.cu

²Instituto de Biotecnología de las Plantas, Universidad Central Marta Abreu de Las Villas. Carretera a Camajuaní km 5.5. Santa Clara. Villa Clara. Cuba. CP 54 830

RESUMEN

Los procedimientos de comparaciones múltiples permiten detectar diferencias entre las medias de los tratamientos. El objetivo de muchos experimentos en Biotecnología Vegetal es detectar relaciones significativas entre diferentes tratamientos. En este trabajo se presentan varias técnicas de comparaciones múltiples y se exponen sus ventajas y limitaciones. Además, se ofrecen algunas recomendaciones para su correcta aplicación.

Palabras clave: ANOVA, comparaciones por pares

ABSTRACT

Multiple comparison procedures enable to detect differences among treatment means. Many plant biotechnological experiments are aimed to detect significant relationships among different treatments. Several multiple comparisons techniques are presented in the current work. Besides, advantages and limitations are exposed. Some recommendations about the correct application are also provided.

Key words: ANOVA, pairwise comparisons

Contenido

INTRODUCCIÓN

ELANALISIS DE VARIANZA

Comparaciones *a posteriori*

Asumiendo varianzas homogéneas

- Método de la mínima diferencia significativa
- Método de Bonferroni
- Método de Sidak
- Método R-E-G-W F : método de Ryan (1960), Einot-Gabriel (1975) y Welsch (1977)
- Método de Student-Neuwman-Keuls: (Newman, 1939; Keuls, 1952)
- Método de Tukey
- Método de Tukey-b
- Método de Scheffé
- Método de Duncan

Asumiendo varianzas diferentes

- Método T2 de Tamhane
- Método T3 de Dunnett
- Método de Games-Howell- Método C de Dunnett

INTRODUCCIÓN

El tratamiento estadístico de los resultados experimentales representa una poderosa herramienta para la interpretación de los datos en cualquier ciencia. En particular, la aplicación de la estadística en las ciencias agrícolas no es un tributo a la modernidad sino una larga tradición que se origina en los trabajos de Sir Ronald Fisher, matemático británico (1890 - 1962). Los proyectos de Fisher, se utilizaron primeramente en biología, pero rápidamente cobraron importancia y se aplicaron a

problemas agronómicos, sentando así las bases de aplicación de la estadística a la experimentación agrícola y biotecnológica.

En este contexto, uno de los problemas, al que se le presta especial atención, es a la comparación de varios grupos o tratamientos. Específicamente, se habla de comparar una variable continua que pudiera ser el rendimiento, o la altura del tallo, con respecto a varias categorías de una variable discreta. Cada una de estas categorías representa un grupo o tratamiento diferente. A esta variable categórica se le denomina factor.

Si el factor tiene solamente dos valores posibles el problema es muy sencillo de resolver desde el punto de vista estadístico. Para este caso, las pruebas clásicas de *Student* y de *Fisher* (Milton, 1999) para la comparación de medias y varianzas suponiendo distribuciones normales pueden utilizarse con facilidad y se encuentran implementados en cualquier paquete de software estadístico (Meulman y Heiser, 2004). Si la normalidad de la variable continua para alguna de las categorías del factor falla, entonces, deben utilizarse métodos no paramétricos que también se hallan en la literatura especializada (Siegel, 1987; Glantz, 2002) y en paquetes de software con cierta facilidad (Meulman y Heiser, 2004).

El problema que se abordará en el presente trabajo surge cuando la cantidad de valores del factor es al menos tres. Cuando este es el caso, pudiera pensarse en la realización de tantas comparaciones por pares como fuera necesario con ayuda de la prueba clásica de *Student*.

Si se imagina una situación particular en la que se tienen tres grupos y se quieren detectar las diferencias entre ellos deben realizarse tres comparaciones por pares: el grupo 1 con el 2, el 1 con el 3 y finalmente el 2 con el 3. Si se supone que cada una de esas comparaciones usa una probabilidad de 0.05 para rechazar hipótesis nulas correctas, (esta probabilidad se conoce como error de tipo I. (Glantz, 2002). Asumiendo que las comparaciones son independientes, la probabilidad general de no cometer un error de tipo I se calcula de la forma: $(0.95)^3 = 0.857$. Luego la probabilidad de cometerlo será: $1 - (0.95)^3 = 0.143$. Nótese que en estas comparaciones grupales la probabilidad de rechazar hipótesis nulas correctas se incrementó de un 5.0% a un 14.3%, considerando solamente tres grupos. Si los tratamientos a comparar fueran cinco entonces $1 - (0.95)^5 = 0.226$ y si fueran ocho $1 - (0.95)^8 = 0.336$ (Field, 2003). Como puede apreciarse en la medida en que se incrementa el número de grupos a considerar aumenta de manera notable la probabilidad de cometer un error de tipo I.

El problema presentado tiene solución desde el punto de vista estadístico, sólo que esta no es trivial. Al método general de comparación de tres o más grupos o tratamientos de manera conjunta se le denomina 'análisis de varianza' (Milton, 1999; Montgomery, 2004; Hall, 2004). Es por ello, que el presente trabajo tuvo como objetivo analizar varias técnicas de comparación múltiples así como realizar recomendaciones acerca de cuando resulta adecuado utilizarlas.

EL ANÁLISIS DE VARIANZA

Para la solución de un problema de investigación o en el campo de la mejora genética con el empleo de

las técnicas de cultivo de tejidos de plantas, surge la necesidad de realizar alguna comparación de tratamientos. Esto permitiría elegir la mejor alternativa de las varias que existen, o por lo menos para tener una mejor comprensión del comportamiento de la variable de interés en cada uno de los distintos tratamientos. Por ejemplo, la comparación de cuatro concentraciones de reguladores del crecimiento (Ej. 6-bencilaminopurina) en la regeneración *in vitro* de brotes de *Phaseolus vulgaris* L., lo cual se hace con el fin de determinar con cuál de las concentraciones se logró la regeneración de plantas en dos genotipos de frijol (Delgado-Sánchez *et al.*, 2006).

El propósito fundamental de un análisis de varianzas (ANOVA) es comparar los grupos o tratamientos entre sí (Milton 1999; Field, 2003; Montgomery, 2004; Hall, 2004). La hipótesis nula o fundamental puede enunciarse en los términos: 'no existen diferencias significativas entre las medias poblacionales de todos los tratamientos'. Si se cumpliera tal hipótesis entonces se puede afirmar que los grupos no se diferencian entre sí y que el factor es independiente de la variable analizada. La hipótesis alternativa por su parte enuncia que existen diferencias significativas entre los tratamientos, es decir que al menos dos de los grupos difieren.

Para probar la hipótesis de igualdad de medias se calcula un estadístico F de Fisher, que refleja el grado de parecido existente entre las medias que se están comparando (Milton 1996; Field, 2003; Montgomery, 2004; Hall, 2004). Si no existen diferencias significativas entre las medias poblacionales, las medias muestrales serán parecidas, existiendo entre ellas tan sólo diferencias atribuibles al azar (Field, 2003).

El estadístico F del ANOVA se basa en el cumplimiento de dos supuestos fundamentales: normalidad y homocedasticidad. Algunos autores coinciden en afirmar que pueden confiarse en los resultados aún si las variables se alejan un poco de la distribución normal y tienen varianzas moderadamente diferentes (Aron y Aron, 2002).

La razón F en un análisis de varianza solo muestra información acerca de si la varianza contenida en los datos se debe a la casualidad o a la influencia de factores externos. Un resultado estadísticamente significativo no ofrece información acerca de dónde están tales diferencias. Es por ello que, como complemento del análisis de varianza, se deben realizar otras pruebas estadísticas que ayuden al investigador a encontrar cuáles grupos difieren (Milton 1996; Glantz, 2003; Field, 2003).

Una vía para lograr este objetivo es realizar múltiples comparaciones entre todos los grupos. A ellas se les llama comparaciones *a posteriori*.

Comparaciones a posteriori

Existe una gran cantidad de métodos de comparación múltiple. Resulta indiscutible que algunos son mejores que otros y también que algunos han sido mucho más utilizados que otros. Entre los numerosos resultados publicados en cultivo de tejidos, fisiología vegetal y ramas afines, se nota un cierto abuso de algunas de estas técnicas y muy poco uso de otras que pudieran conducir a conclusiones mucho más certeras.

Como ya se había enunciado, uno de los prerrequisitos del análisis de varianza es la homocedasticidad. Varias pruebas de comparaciones múltiples son sensibles a la igualdad o no de las varianzas, es por ello que se realizará un análisis dividido a partir de ahora (Meulemn y Heiser, 2004; SPSS, 2005).

Asumiendo varianzas homogéneas

- Método de la mínima diferencia significativa:

Se basa en la distribución t de Student. Fue propuesto por Fisher en 1935. No ejerce ningún control sobre la tasa de error, es decir, cada comparación se lleva a cabo utilizando el nivel de significación establecido (generalmente 0.05), por lo que la tasa de error para el conjunto de comparaciones puede llegar a $1 - (1 - \alpha)^k$, siendo α el nivel de significación y k el número de comparaciones realizadas. Los pares de medias que se comparan (utilizando el estadístico t de Student) son aquellos que han sido planeados antes de ejecutar el experimento, por eso se dice que es una prueba para comparaciones planeadas. Esta prueba puede dar resultados falsamente significativos si el experimentador se dedica a hacer comparaciones entre tratamientos de resultados extremos (Dean y Voss, 1999; Atil y Unver, 2001), por lo que no se recomienda su uso. Esta prueba suele encontrarse en la literatura estadística con su nombre en inglés: LSD = *Least Significant Difference*.

- **Método de Bonferroni:** Se basa en la distribución t de Student y en la desigualdad de Bonferroni. Este método intenta resolver el problema que tiene la aplicación de numerosas pruebas de Student reduciendo la probabilidad de cometer un error de tipo I en cada comparación. Controla la tasa de error dividiendo el nivel de significación α entre el número de comparaciones k realizadas. Cada comparación se evalúa utilizando un nivel de significación $\alpha_B = \alpha/k$. Este procedimiento trabaja razonablemente bien cuando se tiene un número pequeño de grupos (Glantz, 2003). Si la cantidad de tratamientos se incrementa a ocho o a diez, el valor de t requerido para detectar una diferencia será demasiado grande, (el valor de α_B será demasiado pequeño), de manera que en la práctica el método se vuelve incapaz de detectarlas. En tales casos no se recomienda su

utilización. Esta técnica también se conoce con el nombre de Dunn – Bonferroni y se publicó por primera vez en 1961.

- **Método de Sidak:** Se basa en la distribución t de Student, pero controla la tasa de error evaluando cada comparación con un nivel de significación $\alpha_s = 1 - (1 - \alpha)^{1/k}$. Esta solución es algo menos conservadora que la de Bonferroni, es decir, rechaza la hipótesis fundamental en más ocasiones que el método de Bonferroni, pero de manera general es también un método recomendable. Se publicó por primera vez en 1967.

- Método R-E-G-W F : método de Ryan (1960),

Eino-Gabriel (1975) y Welsch (1977): Se basa en la distribución F . Es un método por pasos. Comienza ordenando de forma ascendente los p grupos por el tamaño de sus medias, se efectúan todas las comparaciones posibles entre pares de medias teniendo en cuenta el número de escalones (r) que las separan. Con p medias, la media más pequeña y la más grande están separadas $r = p$ escalones; la media más pequeña y la segunda más grande están separadas $r = p - 1$ escalones; la media más pequeña y la tercera más grande están separadas $r = p - 2$ escalones y así sucesivamente. Después de ordenadas, dos medias adyacentes están separadas por dos escalones. El número de escalones existente entre las medias comparadas condiciona el nivel de significación de cada comparación, siendo éste mayor cuanto más alejadas se encuentran las medias ordenadas. En el método R-E-G-W F , cada comparación se evalúa utilizando un estadístico F y un nivel de significación $\alpha_R = 1 - (1 - \alpha)^{1/p}$. Es un método mucho más potente que el de Duncan y el de Student-Newman-Keuls (ver más abajo), pero no es apropiado cuando los grupos tienen tamaños distintos (Meuleman y Heiser, 2004).

- Método de Student-Newman-Keuls:

(Newman, 1939; Keuls, 1952): Se basa en la distribución del rango *estudentizado*. Al igual que el métodos R-E-G-W F , parte de ordenar las medias por su tamaño. Pero a diferencia de él, aquí el nivel de significación para cada conjunto de medias separadas r pasos es siempre α . Cuantos más pasos existen entre dos medias, mayor es la diferencia mínima necesaria para considerar que esas medias difieren significativamente (Atil y Uver, 2001).

- **Método de Tukey:** Se le conoce como método de la diferencia honestamente significativa de Tukey. Equivale a utilizar el método de Student-Newman-Keuls con $r = p$, por tanto, todas las comparaciones se refieren a una misma diferencia mínima, o sea utiliza un sólo valor con el cual se comparan todos los posibles pares de medias. Numerosos autores coinciden en afirmar que la prueba es apropiada para realizar comparaciones por pares, por lo que tiene

una gran aceptación (Atil y Uver, 2001). Se publicó por primera vez en 1953.

- **Método de Tukey-b:** Es una modificación del anterior. Consiste en considerar como diferencia mínima el valor medio entre la diferencia honestamente significativa de Tukey y la diferencia mínima obtenida con el método de Student-Newman-Keuls para el caso de $r = 2$. Se publicó por primera vez en 1953 (Dean y Voss, 1999).

- **Método de Scheffé:** Se basa en la distribución F de Fisher. Permite controlar la tasa de error para el conjunto total de comparaciones que es posible diseñar con p medias. Utilizado para efectuar sólo comparaciones por pares, es un procedimiento conservador: tiende a considerar significativas menos diferencias de las que debería. Para ser consecuente con este problema, Scheffé recomienda utilizar $\alpha = 0.10$ (Scheffé, 1959). La prueba se equivoca muy poco en las diferencias que detecta, por lo que su uso es ampliamente recomendado (Atil y Uver, 2001). Si el número de réplicas es diferente en los tratamientos o grupos a comparar, el método de Scheffé resulta ser más apropiado que el de Tukey. Se publicó por primera vez en 1953.

- **Método de Duncan:** Es conocido como la prueba de los rangos múltiples de Duncan. Es un método de comparación por pasos. Controla la tasa de error utilizando, para el conjunto de medias separadas en r pasos, un nivel de significación $\alpha_p = 1 - (1 - \alpha)^{r-1}$. Mientras más pasos existan entre dos medias, mayor es la diferencia mínima con la que se va a considerar que esas medias difieren significativamente. El método es más potente que el de Student-Newman-Keuls, pero no protege adecuadamente el error de tipo I. Tampoco resulta confiable su aplicación cuando el número de réplicas no es igual en los tratamientos que se comparan, por lo que de manera general no se recomienda su uso. Se publicó por primera vez en 1955.

- **Método de Dunnett:** Es una prueba para realizar comparaciones planeadas. Se utiliza fundamentalmente para comparar cada grupo con un grupo control. Este procedimiento permite detectar todos los tratamientos que son tan buenos como el control o mejores que él. Un ejemplo de su aplicación puede ser en estudios toxicológicos en los que el investigador quiere determinar la dosis más baja en la que aún existe una respuesta observable (Atil y Uver, 2001). Se supone que todos los tratamientos tengan el mismo número de casos y controla la tasa de error para $k - 1$ comparaciones. Se publicó por primera vez en 1955.

Asumiendo varianzas diferentes

En el caso en que las varianzas poblacionales sean diferentes deberá utilizarse uno de las siguientes métodos para realizar las comparaciones múltiples:

- **Método T2 de Tamhane:** es un método basado en la distribución del módulo máximo estudentizado. Se publicó por primera vez en 1977.

- **Método T3 de Dunnett:** este método es una modificación propuesta por Dunnett al estadístico T2 de Tamhane. Se basa también en la distribución del módulo máximo *estudentizado* y fue publicado en 1980.

- **Método de Games-Howell:** es un método muy similar al de Tukey. Se basa en la distribución del rango *estudentizado* y en un estadístico t en el que, tras estimar las varianzas poblacionales suponiendo que son distintas, se corrigen los grados de libertad mediante la ecuación de Welch. En términos generales, de los cuatro métodos de este apartado, el de Games-Howell es el que mejor controla la tasa de error en diferentes situaciones. Fue publicado en 1976.

- **Método C de Dunnett:** método idéntico al de Games-Howell excepto en la forma de corregir los grados de libertad de la distribución del rango *estudentizado*. Esta solución es más conservadora que la de Games-Howell. Se publicó en 1980.

Existe una gran cantidad de métodos para comparar tratamientos. Es tarea del investigador seleccionar el o los adecuados, aunque no siempre esta es una tarea sencilla.

Varios autores coinciden en afirmar que la prueba de Scheffé puede aplicarse en cualquier situación, sin embargo, cuando la cantidad de réplicas en los tratamientos es similar, se ha demostrado que el método de Bonferroni es más potente (Atil y Uver, 2001). Otros autores afirman que es la prueba de Tukey la que ofrece resultados más confiables con mayor frecuencia y por tanto, esta es la que debe usarse cuando el investigador se encuentra totalmente desorientado.

En una comparación entre varios métodos se llegó a la conclusión de que si el número de réplicas es diferente en cada grupo, la prueba de Scheffé es mejor que la de Tukey, pero si el diseño es equilibrado, es mejor utilizar la prueba de Tukey en vez del de Scheffé (Atil y Uver, 2001).

La prueba de Duncan por su parte es un método de uso muy común, al igual que la prueba de la mínima diferencia significativa, debido a la simplicidad de sus cálculos. Existen paquetes estadísticos profesionales que tienen estas técnicas implementadas, luego la dificultad en los cálculos ya no es un problema real. Ambos métodos incrementan de manera notable el error de tipo I, por lo que no se recomienda su uso en ninguna situación (Aron y Aron, 2002; Dean y Voss, 1999; Atil y Uver, 2001).

Como puede apreciarse no existe un único método que ofrezca siempre resultados correctos. La elección del método de comparación adecuado es una tarea difícil que debe resolver el investigador, en dependencia de las características de su experimento y de los objetivos de su investigación. En este trabajo se presentan algunas recomendaciones que deben tenerse en cuenta a la hora de seleccionar el método apropiado.

REFERENCIAS

- Aron A, Aron E (2002) Statistics for the Behavioral and Social Sciences. A Brief Course. Second Edition. Vol. 1. p.330. Prentice Hall. New York
- Atil H, Unver Y (2001) Multiple Comparisons. OnLine Journal of Biological Sciences 1(8): 723-727
- Dean A, Voss D (1999) Design and Analysis of Experiments. Springer Texts in Statistics. Springer. Dordrech
- Delgado-Sánchez P, Saucedo-Ruiz M, Guzmán-Maldonado SH, Villordo-Pineda E, González-Chavira M, Fraire-Velázquez S, Acosta-Gallegos JA, Mora-Avilés A (2006) An organogenic plant regeneration system for common bean (*Phaseolus vulgaris* L.). Plant Science 170: 822–827
- Einot, I, Gabriel, KR (1975) A study of the power of several methods of multiple comparisons. Journal of the American Statistical Association 70: 574-583
- Field, A (2003) Discovering Statistics Using SPSS for Windows. Advanced Techniques for the Beginner. Vol. 1 p. 496. Daniel B. Wright
- Glantz, S (2002) Primer of biostatistics. Fifth edition. Vol 1. p. 468. McGraw-Hill. New York
- Hall, C (2004) A Handbook of Statistical analysis Using SPSS
- Keuls, M (1952) The use of the Studentized range in connection with and analysis of variance. Euphytica 1: 112-122
- Meulman J, Heiser WJ (2004) Manual de SPSS
- Milton, S (1999) Statistical Methods in the Biological and Health Sciences. Third Edition Statistics. Vol. 1, p. 586. McGraw-Hill. New York
- Montgomery, D C (2004) Diseño y análisis de experimentos. Tercera edición. Editorial Félix Varela. La Habana
- Newman, D (1939) The distribution of the range in samples from a normal population expressed in terms of an independent estimate of standard desviation. Biometrika 31: 20-30
- Ryan, TA (1959) Multiple comparison in psychological research. Psychological Bulletin 56: 26-47
- Scheffé, H (1959) The Analysis of Variance. Publications in Statistics. John Wiley and Sons. Wiley. New York
- Siegel, S (1987) Diseño experimental no paramétrico. Edición Revolucionaria p.335. La Habana
- SPSS (2005) Manual del SPSS p.13.
- Welsch, RE (1977) Stepwise multiple comparison procedure. Journal of the American Statistical Association 72: 566-575