Psicothema
2000. Vol. 12, Supl. nº 2, pp. 114-116

ISSN 0214 - 9915 CODEN PSOTEG

Estudio de la potencia de los contrastes de medias con dos y tres grupos con tamaño de efecto pequeño y en condiciones de no normalidad y homo-heterocedasticidad

Isabel Cañadas Osinski, África Borges del Rosal, Alfonso Sánchez Bruno y Concepción San Luis Costas
Universidad de La Laguna

Se estudia el problema de la potencia analizando tamaños de efecto pequeño en dos y tres muestras, utilizando diversos contrastes paramétricos y no paramétricos (*t* de Student, prueba de Welch y U de Mann-Withney-Wilcoxon en el caso de dos grupos, y ANOVA, prueba de James de segundo orden (James, 1951, 1954) y Kruskal-Wallis, en el caso de tres grupos), ello tanto en condiciones de cumplimiento de los supuestos subyacentes a las pruebas como bajo violación de los supuestos de normalidad y homocedasticidad. La principal conclusión es que, cuando no sea posible trabajar con muestras numerosas, (caso común en la investigación en Psicología aplicada) se deben plantear otras formas de investigación que permitan resultados más esclarecedores.

The power of mean tests with two and three groups, with small effect size and under non-normality and homo-heteroscedaticity. The power of mean contrasts is studied in conditions of small effect size, with two and three samples, and using several parametric and non-parametric statistical tests (Student's test, Welch test and Mann-Withney-Wilcoxon's U with two groups; ANOVA, James second order test and Kruskal-Wallis with three groups), both complying with the parametric assumptions of normality and homocedasticity and violating those assumptions. The main result is that, when it is not possible to have big samples (which is very common when investigating in applied Psychology) these tests are not useful and different ways of investigating must be accomplished.

La potencia de los contrastes estadísticos ha sido un tema tradicionalmente poco tenido en cuenta por los investigadores aplicados y no muy estudiado por los metodólogos. En un estudio pionero sobre este tema, Cohen (1962) encontró que en la inmensa mayoría de los trabajos publicados en las revistas científicas de Psicología la potencia de los contrastes era extraordinariamente baja. Sedlmeier y Gigerenzer (1989) o Clark-Carter (1997) encontraron que, tras veinte años en el primer caso y treinta en el segundo, la situación de los estudios en lo que respecta a la potencia se encontraba exactamente igual. Sedlmeier y Gigerenzer (1989) comentan:

> «Como resultado general, por lo tanto, la potencia no se ha incrementado después de 24 años. Aunque ahora hay un pequeño porcentaje de experimentos en los cuales la posibilidad de encontrar un resultado significativo si hubiera efecto, es alta, incluso para efectos pequeños o medianos, la potencia mediana respectiva para efectos medios es un poco menor que lo que se encontró hace 24 años».

Correspondencia: Isabel Cañadas Osinski
Facultad de Psicología
Universidad de La Laguna
38200 Tenerife (Spain)
E-mail: Icanadas@ull.es

En una revisión de investigaciones clínicas, Kazdin y Bass (1989), tras revisar 120 estudios publicados en revistas de investigación en psicoterapia entre 1984 y 1986, encontraron una gran despreocupación con respecto al tema de la potencia. Para estos autores es corriente encontrarnos con la afirmación que no hay diferencias entre tratamientos alternativos para una problemática dada, especulándose con la existencia de factores comunes a cualquier orientación terapéutica. Sin embargo, una explicación diferente podría ser que la falta de potencia de los contrastes haya impedido que tales diferencias salgan a la luz, sobre todo cuando los tamaños muestrales utilizados en las investigaciones son demasiado pequeños para que puedan cubrir la potencia mínima que permitiría detectar efectos pequeños o medianos.

Con respecto a los tamaños muestrales, estos mismos autores encontraron que la mediana de los tamaños grupales en las publicaciones estudiadas estaba en 12, con un rango de 3 a 114 y que tres cuartas partes de los estudios incluían menos de 20 sujetos por grupo. En otro estudio sobre este mismo tema, Shapiro y Shapiro (1982) encontraron que los tamaños grupales incluidos en los estudios eran en un 10% de los trabajos, 6 o menos sujetos por grupo de tratamiento, en un 26% de 7 a 9, en un 36% de 10 a 12 y sólo en un 28% se incluían más de 13 sujetos.

Para Keppel (1991) existe un cierto consenso en el sentido de que se debe ir a potencias de alrededor de 0,8, lo que representa un criterio razonable y realista para las ciencias del comportamiento, en el sentido de que con una relación 4:1 refleja el sentimiento general de que el error de tipo I es más grave y al mismo tiempo pro-

tege razonablemente contra errores de tipo II. Pese a ello, la realidad es que a menudo aspirar a tales niveles de potencia nos condena a detectar únicamente tamaños de efecto muy altos o a trabajar con tamaños muestrales inabarcables fuera del ámbito de la simulación.

Estas consideraciones nos han llevado a estudiar el problema de la potencia, analizando tamaños de efecto pequeño en dos y tres muestras, utilizando diversos contrastes paramétricos y no paramétricos (*t* de Student, prueba de Welch y U de Mann-Withney-Wilcoxon en el caso de dos grupos, y ANOVA, prueba de James de segundo orden (James, 1951, 1954) y Kruskal-Wallis, en el caso de tres grupos), ello tanto en condiciones de cumplimiento de los supuestos subyacentes a las pruebas como bajo violación de los supuestos de normalidad y homocedasticidad.

Método

El procedimiento seguido ha sido el de simulación de muestras, realizando 185.000 replicaciones.

Se ha estudiado la potencia del contraste, estimada mediante la proporción de rechazos de la hipótesis nula, con tamaño de efecto pequeño en todos los casos. Para la definición del tamaño de efecto, d, se siguió a Cohen (1988), de forma que el valor de d es 0.2 y se añade a $\bar{x}_2$ en todos los casos.

Como variables independientes se consideraron la homo-heterocedasticidad, con valores de la varianza 1 y 3 para el caso de dos grupos y 1, 2, y 3 para el caso de 3 grupos; el tamaño de grupo (5, 10, 15, 20, 25 y 30 sujetos simulados) con sus diferentes combinaciones; y la forma de la distribución, cuantificada mediante los coeficientes de asimetría y apuntamiento:

Asimetría, con los siguientes valores: -2, -1, 0, 1 y 2, en el caso de dos grupos, y normalidad, -1 y 1 en tres grupos. Es importante destacar que el procedimiento de Fleishman, utilizado por nosotros, no permite generar distribuciones con tales niveles de asimetría en ausencia de apuntamiento, por lo que todas las distribuciones se generaron con apuntamiento de 5,2.

Apuntamiento, con los siguientes niveles: normal (0), leptocúrtico (5,2) y platicúrtico (-1,15), todos ellos con asimetría cero.

Resultados

Como resultado general debemos decir que la potencia alcanzada, en todos los casos, es despreciable. Ahora bien, incluso en las condiciones óptimas (distribuciones normales y homocedásticas), las potencias esperadas van desde 0,06 cuando los grupos tienen un tamaño de 5, hasta 0,12 en el caso mejor, esto es, cuando los grupos tienen 30 observaciones. Estos valores están muy alejados, obviamente, de los aconsejables, según el consenso que señalaba Keppel (1991), de 0,8.

I) Dos muestras

a) Homocedasticidad

1) Apuntamiento

Según se desprende de nuestros resultados, cuando existe homocedasticidad las violaciones de la normalidad por diferencias en el apuntamiento no provocan problemas importantes en la potencia del contraste t de Student que, de hecho, puede incluso incrementarse cuando una o ambas distribuciones son leptocúrticas. Los mínimos problemas que se pueden presentar cuando alguna de las distribuciones es platicúrtica se solucionan fácilmente trabajando con grupos de tamaño 10 o superior y, preferiblemente, homogéneos.

En cuanto a la prueba de elección en estas circunstancias, existen casos concretos en los que los contrastes alternativos se comportan mejor que la t de Student, pero como en otras circunstancias su comportamiento es mucho peor, creemos que debe ser ésta la prueba escogida en caso de violaciones del apuntamiento.

2) Asimetría

En condiciones de homocedasticidad, la asimetría no es un problema para la t de Student cuando se produce en el mismo sentido en ambas distribuciones pudiendo, incluso, mejorar la potencia de los contrastes con respecto a la condición de normalidad en ambas distribuciones.

Por otra parte, en trabajos anteriores (Cañadas y cols. en prensa) vimos que en el caso de asimetrías cruzadas falla la robustez de todos los contrastes estudiados; ahora añadimos, además, que si la distribución con media menor es asimétrica negativa y la distribución con media mayor asimétrica positiva, falla también la potencia, no existiendo en estos casos una solución clara basada en estos contrastes. Lo recomendable posiblemente sea modificar la estrategia global de la investigación, pero si es imprescindible utilizar un contraste de medias, lo único que podemos recomendar es trabajar con la t de Student y tomar muestras del mayor tamaño posible.

b) Heterocedasticidad

En el trabajo anterior ya citado (Cañadas y cols, en prensa), habíamos comprobado que sólo el contraste de Welch se mantenía robusto con tamaños grupales distintos. Ahora bien, puesto que este contraste presenta resultados desalentadores cuando se estudia su potencia, una vez más nos vemos incapacitados para hacer una recomendación que no sea la de cambiar la estrategia global de la investigación.

II) Tres muestras

a) Homocedasticidad

1) Apuntamiento

Los tres contrastes se comportan de forma similar en las distintas condiciones de apuntamiento, sin diferenciarse de los resultados que aparecen en condiciones de normalidad. El contraste Kruskal-Wallis cuando dos de las distribuciones, o las tres, son leptocúrticas, es el que muestra valores de potencia más altos (excepción hecha de cuando dos o tres de los grupos tienen tamaño 5).

Sin embargo, el comportamiento de los contrastes cambia drásticamente cuando la aumenta variabilidad, incluso si se mantiene constante entre los tres grupos. De los tres contrastes, el de James es el que presenta mayores problemas, llegando incluso a valores de 0,06 en los tamaños grupales mayores, sin que las diversas condiciones de forma de las distribuciones tengan una contribución específica en el comportamiento de la prueba.

2) Asimetría

En lo que respecta a asimetría, el Anova y la prueba de James tienen un comportamiento similar, mientras que el contraste de Kruskal-Wallis es consistentemente más potente. No obstante, debido a la ausencia de robustez cuando las distribuciones tienen distintas asimetrías (Borges y cols, en prensa), sólo sería aconsejable frente al Anova cuando, existiendo gran variabilidad, pero constante entre los grupos, éstos presentaran la misma forma.

b) Heterocedasticidad

Como ya apuntábamos en el apartado de homocedasticidad, la potencia parece verse más afectada por el aumento de la variabilidad de las muestras que por la desigualdad de las varianzas. Ineludiblemente, sea cual sea el contraste utilizado y la forma de la distribución, el aumento en la variabilidad de las distribuciones comporta pérdida de potencia. De los tres contrastes, el más afectado es el James, que muestra valores de potencia francamente despreciables.

Conclusiones

La primera conclusión que cabe destacar es que si el investigador pretende detectar un efecto pequeño y el tamaño de sus grupos es escaso, tendrá que plantearse una forma de investigación distinta a la que aquí hemos analizado, puesto que alcanzar un resultado significativo es altamente improbable.

Los problemas prácticos que se derivan de los resultados obtenidos son relevantes. En psicología, las características de las investigaciones no permiten esperar la aparición de efectos grandes, sino en todo caso medianos o, tal vez, pequeños (Clark-Carter, 1997). Teniendo en cuenta que la ausencia de resultados significativos supone problemas como la imposibilidad de publicar los trabajos y/o que, a la postre, se llegue al abandono de la línea de trabajo, el abordar la investigación desde el contraste de significación y con planteamientos clásicos, conducirá, inevitablemente, a producir pocos avances de interés.

Creemos, por tanto, a la luz de los resultados obtenidos, que cuando las poblaciones de estudio no permitan grupos numerosos (caso común en la investigación en Psicología aplicada) se deben plantear otras formas de investigación que permitan resultados más esclarecedores.

Si pretendemos seguir utilizando la estadística, podemos, según Cohen (1990), o bien quedarnos en el nivel descriptivo o, si deseamos utilizar la inferencia, sería más aconsejable informar del tamaño del efecto obtenido que conformarnos con la información SI/NO que facilita el contraste de significación.

Por otro lado, cabe plantearse otra forma de investigación; así, por ejemplo, los diseños de caso único pueden ser la alternativa cuando se trata de analizar cambios debidos a la intervención psicológica y contamos con un escaso número de sujetos. La investigación será un poco más complicada que el mero análisis de dos conjuntos de datos (si utilizamos el clásico y poco interesante diseño *grupo tratamiento / grupo control*) pero, definitivamente, puede resultar mucho más fructífera.

Nuestros resultados ponen en evidencia que la polémica suscitada en torno a la utilidad del contraste de significación, no sólo mantiene total vigencia, sino que exige que los metodólogos alertemos a los investigadores aplicados sobre los peligros de utilizar instrumentos estadísticos que no se ajustan a las condiciones de los datos, por una parte, y a ofertar procedimientos de investigación más fructíferos, por otra.

Referencias

Borges, A., Cañadas, I. y Sánchez-Bruno, A. (en prensa). El contraste de hipótesis en tres grupos: alternativas al anova frente a la violacion de sus supuestos. *Psicothema.*

Cañadas, I., Borges, A. y Sánchez-Bruno, A. (en prensa). La t de Student y sus alternativas, ante la violación de los supuestos. *Psicothema.*

Clark-Carter, D. (1997) The account taken of statistical power in research published in the British Journal of Psychology. *British Journal of Psychology, 88*, 71-83.

Cohen, J. (1962) The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

Cohen, J. (1990) Things I have learned (so far). *American Psychologist, 45*, 1.304-1.312.

James, G.S. (1951) The comparison of several groups of observations when the ratios of the population variances are unkonown. *Biometrika, 38*, 324-329.

James, G.S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unkonown. *Biometrika, 41*, 19-43.

Kazdin, A.E. y Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57*, 138-147.

Keppel (1991). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice Hall.

Sedlmeier, P. y Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies?. *Psychological Bulletin, 105*, 309-316.