Correcciones Bonferroni y Sidák para los Efectos de la Multiplicidad con Puntajes Subtotales en Pruebas Poligráficas con Preguntas de Comparación

Raymond Nelson

Abstract

Se discute el problema de las comparaciones estadísticas múltiples en cuanto a cómo se aplica en el uso de puntajes subtotales de pruebas poligráficas con pregunta de comparación. El fenómeno de la multiplicidad incluye la inflación del alfa cuando se utiliza cualquiera de un conjunto de puntajes subtotales múltiples para hacer clasificaciones de engaño, y la deflación del alfa cuando se usa todo un conjunto de subtotales múltiples para hacer clasificaciones de veracidad de los resultados de prueba. Se describen correcciones estadísticas comunes, que incluyen la corrección de Bonferroni y la corrección de Šidák. Se proporcionan ejemplos matemáticos para ilustrar la aplicación de estas correcciones estadísticas en la prueba de polígrafo de preguntas de comparación.

Palabras Clave: *puntuación, análisis de datos de prueba, técnicas poligráficas, data normativa, distribuciones de referencia, significancia estadística, Bonferroni, Sidák, comparaciones múltiples, alfa inflado, alfa desinflado, multiplicidad*

Traductores:
Juan Carlos Padilla Sanabria
jpadilla@deltasis.com.co

Rodolfo Prado Pelayo
rodolfo@poligrafia.com.mx

Introducción

Los efectos de la multiplicidad, también conocidos como el problema de las comparaciones múltiples (McDonald, 1996; Miller, 1981), son bien conocidos por los científicos, investigadores, estadistas y otros profesionales cuyo trabajo involucra la evaluación de datos como una base para la clasificación y la inferencia. Estos efectos también se han referido como el efecto "mirar a otro lugar" (White, 2011), debido al impulso o deseo de seguir buscando en otro lugar cuando no encontramos inicialmente lo que estamos buscando. En el contexto de la investigación y de las pruebas científicas, los efectos de la multiplicidad y el impulso de seguir buscando en otro lugar hasta que encontramos lo que estamos buscando, puede pensarse como una manifestación de la confirmación de un prejuicio descrito por Nickerson (1998). Cuando no estamos satisfechos ignoramos los resultados y continuamos buscando hasta que encontramos un resultado con el cual estamos satisfechos.

Una analogía del juego de cartas puede ser útil para entender mejor las implicaciones prácticas: imagina a un jugador de póker que se reparte a sí mismo una mano de cartas con el objetivo de hacerlo repetidamente hasta conseguir una Flor Imperial. La teoría de la probabilidad nos dice que, con un número suficiente de ensayos, las probabilidades se acumularán a un nivel lo suficientemente alto como para que eventualmente podamos observar su ocurrencia. Pero, asumiendo un mazo de cartas justo e imparcial, sería un error tratar de inferir que el mazo de cartas tiene características especiales o que el jugador tiene atributos únicos que causaron la Flor Imperial. En su lugar, la ocurrencia de la Flor Imperial simplemente está en función de continuar buscando en otro lugar (en manos posteriores de las cartas) para su aparición. Del mismo modo, mirar repetidamente cualquier conjunto de datos científicos puede confundir nuestros intentos de hacer inferencias realistas y precisas acerca de la significancia o el significado cuando eventualmente observamos lo que estamos buscando. Más específicamente, los efectos de la multiplicidad son la combinación de probabilidades de error. Ellos pueden dar como resultado una pérdida de la exactitud o precisión y el correspondiente aumento en el error de clasificación.

Discusión

Los efectos de la multiplicidad juegan un rol en los exámenes poligráficos con preguntas de comparación cuando se utilizan puntajes subtotales para clasificar los resultados como de engaño o veracidad. Los puntajes subtotales para preguntas relevantes individuales han demostrado ser una base efectiva para las clasificaciones de engaño cuando la puntuación de gran total no es concluyente (Senter, 2002; Senter & Dollins, 2003; 2008). Sin embargo, las técnicas poligráficas que hacen uso de los puntajes de gran total, de manera consistente han producido tasas de precisión más altas que las técnicas para las cuales las decisiones se basan únicamente en los puntajes subtotales (APA, 2011). Los puntajes subtotales han sido la base tradicional con los que se clasifican los resultados de los exámenes exploratorios de temas múltiples (Department of Defense, 2006a, 2006b) cuando se puntúa manualmente. Las puntuaciones manuales de gran total tradicionalmente no se usan en las pruebas exploratorias de asuntos múltiples.

El uso de puntajes poligráficos subtotales como base para la clasificación e inferencia estadística introducirá incrementos en la probabilidad de errores matemáticos y estadísticos conocidos y predecibles, a menos que se apliquen correcciones. Estos efectos ocurren porque cada resultado de prueba imperfecto y no determinista es un resultado probabilístico. Siempre hay alguna probabilidad asociada de que el resultado sea correcto e incorrecto. La estimación del error de prueba será una suma de los errores para todos los resultados probabilísticos usados para clasificar el resultado de la prueba.

Inflación del alfa para resultados de engaño de los polígrafos diagnósticos de evento especifico.

El uso de puntajes subtotales en exámenes de evento específico, para los cuales se hará una clasificación al nivel de la prueba como un todo, introduce multiplicidad dentro del modelo estadístico. En la práctica esto equivale a tomar decisiones estadísticas múltiples con respecto a una sola clasificación. Cuando se realizan juicios probabilísticos múltiples con respecto a un incidente o alegato de un solo objetivo, para el que cualquier resultado subtotal de engaño dará como resultado una clasificación del examen como de engaño, la probabilidad de error resultante es la probabilidad acumulada o adicionada de la probabilidad error de todos los puntajes subtotales. En el caso de un polígrafo diagnóstico de evento específico con tres preguntas relevantes (RQ) y alfa = .05, la probabilidad de error total puede determinarse sumando los niveles alfa para todas las RQs (.05 + .05 + .05 - .15). Los cálculos indican una tasa potencial de error del 15% a pesar de que la prueba se realice con un alfa= .05, con el objetivo de limitar los errores a una tasa inferior al 5%. En ocasiones esto se ha referido como el problema de *alfa inflado* debido al aumento predecible en los errores de la prueba. Dejar el alfa inflado no administrado, puede dar lugar a una tasa de error falso positivo que es potencialmente varias veces mayor de lo que fue previsto o anticipado. Por lo tanto, aunque el objetivo era limitar los errores falsos positivos al 5%, la práctica de utilizar subtotales aumenta esa tasa de error falso positivo a aproximadamente 15%.

Corrección Bonferroni.

Afortunadamente, el problema del alfa inflado es levemente irritante y relativamente fácil de rectificar mediante el uso de una corrección estadística simple: la *corrección de Bonferroni* (Abdi, 2007), llamada así por el famoso estadista italiano Carlo Emilio Bonferroni (1892-1960). La corrección de Bonferroni se calcula dividiendo el nivel del alfa deseado por el número de decisiones estadísticas. El número de decisiones estadísticas es igual al número de puntajes subtotales, que es igual al número de RQs. El nivel alfa corregido resultante es conocido como *el alfa corregido Bonferroni.*

Para polígrafos diagnósticos de evento específico con tres RQs y un alfa deseado de .05 dividimos .05 entre tres (alfa = .05 / 3 RQs = alfa .0167 por RQ). Será necesario usar el alfa corregido de Bonferroni = .0167 para cada uno de los tres puntajes subtotales. Cuando estas probabilidades de error por pregunta se acumulan (.0167 + .0167 + .0167 = .05) el margen de error acumulado total para la prueba será de un alfa = .05. La estimación del error estará limitada dentro del rango deseado menor al 5%. Los exámenes diagnósticos de evento específico con dos

RQs requerirán el uso del alfa corregido de Bonferroni = .025. Esto es porque el alfa = .05 / 2 RQs = .025 por cada RQ, y esto se acumulará .025 + .025 = .05. Igualmente, los exámenes diagnósticos de evento específico con cuatro RQs usarán una corrección Bonferroni del alfa = .05 / 4 = .0125 por RQ, que se acumulará a .0125 + .0125 + .0125 + .0125 = .05. Debido a que los puntajes subtotales no se utilizan para hacer clasificaciones veraces para los exámenes diagnósticos de evento específico, no se necesita una corrección estadística para la clasificación de veraz en este tipo de exámenes.

Deflación del alfa para resultados veraces en exámenes exploratorios de asuntos múltiples.

Los exámenes exploratorios de asuntos múltiples utilizan los puntajes subtotales para ambas clasificaciones, tanto de engaño como de veracidad. Esto se logra con el criterio de "todos o cualquiera de" que establece que *cualquier* resultado subtotal que sea de engaño será suficiente para clasificar el resultado del examen como de engaño, mientras que *todos* los resultados subtotales deben indicar veracidad para clasificar el resultado general del examen como veraz. Los resultados son inconclusos cuando una o más de las puntuaciones subtotales no son estadísticamente significativas para veracidad y ninguna de las puntuaciones subtotales es estadísticamente significativa para engaño. Al igual que con los exámenes diagnósticos de evento específicos, la estadística de error de prueba para un polígrafo exploratorio de temas múltiples está en función del número de puntajes subtotales (el número de preguntas relevantes). Este fenómeno se aplica a todas las formas de prueba que involucran comparaciones estadísticas múltiples.

Con los exámenes diagnósticos de evento específico, todas las preguntas relevantes describen detalles relacionados con un solo alegato o incidente. Las preguntas relevantes para los polígrafos exploratorios de asuntos múltiples describirán asuntos comportamentales distintos con una fuerte suposición de independencia. La suposición de independencia no se basa únicamente en la premisa del uso de diferentes verbos de acción o del contenido semántico de cada pregunta relevante. Involucra también la suposición de que un examinado podría involucrarse en una o más conductas, mientras que es posible que permanezca completamente sin involucramiento en otras conductas. (Este supuesto de independencia no se utiliza en el polígrafo diagnóstico de evento específico para el cual todas las preguntas relevantes describen aspectos de un solo alegato o incidente.)

Se dice que la suposición de independencia es una fuerte suposición porque, en realidad, aunque se puede asumir que las *conductas* objetivo son independientes o no se afectan entre sí, las *respuestas* del examinado a las preguntas de estímulo del polígrafo de asuntos múltiples *no son* completamente independientes. Las respuestas a diferentes estímulos objetivo pueden afectarse entre sí en un examen. Esto se debe a que todas las respuestas a los estímulos exploratorios de asuntos múltiples tienen una fuente importante de varianza compartida - el examinado. El hecho de que las respuestas no sean completamente independientes parece ser la base de la necesidad del criterio "cualquiera de o todas" y de las prohibiciones tradicionales en contra del

intento de hacer clasificaciones tanto de engaño como de veracidad en un solo examen. Debido a que el criterio cualquiera de o todas no permite resultados veraces y de engaño, esto elimina el potencial de observar tanto errores falso positivo y falso negativo en un solo examen. En cambio, los errores de prueba observados tendrán la forma de errores falso positivo o falso negativo, con lo cual podemos restringir su ocurrencia a los niveles deseados.

Debido a que los asuntos objetivo para los exámenes exploratorios de asuntos múltiples se tratan de manera independiente, no existe una gran preocupación de que estemos sometiendo a un solo asunto objetivo a decisiones estadísticas múltiples. Las pruebas exploratorias pretenden identificar posibles problemas que podrían evaluarse posteriormente a mayor detalle, y, por lo tanto, la sensibilidad de la prueba es una preocupación importante. Las correcciones estadísticas no se utilizan cuando existe una pérdida de sensibilidad potencialmente costosa que reduciría la efectividad de la prueba (McDonald, 2009). Por estas razones, la corrección de Bonferroni no se utiliza para hacer clasificaciones de engaño en los polígrafos exploratorios de asuntos múltiples. Las clasificaciones de engaño para los polígrafos exploratorios de asuntos múltiples se hacen con el límite alfa no corregido.

La multiplicidad juega un rol importante en las clasificaciones *veraces* de los polígrafos exploratorios de asuntos múltiples, pero de una manera ligeramente distinta. Las clasificaciones veraces se hacen cuando los datos observados difieren en un nivel estadísticamente significativo de las distribuciones estadísticas de referencia para los casos de engaño. Por lo tanto, el alfa para clasificaciones veraces representa la tolerancia al riesgo o error de que una persona con engaño podría ser calificada como veraz en una prueba exploratoria de asuntos múltiples (un error falso negativo). Obviamente, se puede esperar que la mayoría de las personas con engaño produzcan puntajes de prueba de engaño, y la proporción de personas con engaño que producen una puntuación de pregunta de prueba que es estadísticamente significativa para veraces (es decir, que difiere en un nivel estadísticamente significativo de las distribuciones de referencia normativas para casos de engaño) se espera que se observe en el nivel definido de alfa (.05). Quizá es igualmente obvio el hecho de que la proporción de personas con engaño que producen *dos* puntajes veraces estadísticamente significativos en una prueba con dos preguntas relevantes será menor que la proporción de personas con engaño que producen una sola puntuación estadísticamente significativa de veracidad. De manera similar, se puede esperar que la proporción de personas con engaño que producen tres de tres puntajes veraces, o cuatro de cuatro puntajes veraces, sea aún menor. Este fenómeno puede pensarse como la *deflación del alfa* que ocurre como resultado del requisito de que el examinado supere todas las preguntas para pasar la prueba. La deflación del alfa dará como resultado una reducción de la tasa observada de error falso negativo a algo predeciblemente menor que la tolerancia alfa establecida para el error.

La deflación de alfa reducirá los errores de prueba para las clasificaciones de engaño, pero también tendrá un efecto en las clasificaciones de veracidad. El requisito de que todos los puntajes subtotales sean estadísticamente significativos para ser

veraces efectivamente proporcionará al examinado veraz la oportunidad múltiple para no producir un puntaje de veracidad estadísticamente significativo. Esta es una característica simple del hecho de que todas las pruebas son probabilísticas y no deterministas, y que las probabilidades pueden ser acumulativas bajo estas circunstancias. Por esta razón, se puede esperar que el requisito de puntajes estadísticamente significativos veraces para *todos* los subtotales provoque una *inflación sustancial de resultados inconclusos* para las personas veraces, junto con una reducción sustancial correspondiente de la especificidad de la prueba para veraces, a menos que se utilice una corrección estadística.

Corrección de Šidák.

La corrección estadística preferida para las clasificaciones de veracidad en los polígrafos exploratorios de asuntos múltiples no es la corrección Bonferroni, sino que es un procedimiento relacionado llamado *corrección de Šidák* (Abdi, 2007, Šidák, 1967). La corrección de Šidák lleva el nombre de Zbyněk Šidák (1933-1999), un reconocido estadista checo. Es una versión exacta de la corrección simple de Bonferroni que se adapta mejor al contexto de las clasificaciones múltiples independientes. Por lo tanto, el cálculo de la corrección de Šidák es: $1-(1-\text{alfa})^{\text{número-de-decisiones}}$. La corrección de Šidák es el complemento matemático del complemento del alfa elevado al número de decisiones. Al igual que con la corrección de Bonferroni previamente descrita, el número de decisiones es igual al número de puntajes subtotales, que también es igual al número de preguntas relevantes.

La forma normal de la corrección de Šidák se usa para calcular la inflación del alfa. Pero nosotros estamos preocupados por la deflación del alfa, por lo que será el inverso de la corrección de Šidák lo que se use para calcular esta deflación. El inverso de la corrección de Šidák se calcula con la siguiente ecuación: $1-(1-\text{alfa})^{1/\text{número-de-decisiones}}$. El inverso de Šidák es el complemento matemático del complemento del alfa elevado al inverso del número de decisiones.

Para demostrar la aplicación de la corrección de Šidák para ajustar o corregir el límite alfa para el número de preguntas relevantes, considere el siguiente ejemplo: un polígrafo de asuntos múltiples con 4 preguntas relevantes para las que el alfa = .05 dará el siguiente nivel alfa no corregido, desinflado: $1-(1-.05)^{1/4} = .0127$. Eso significa que en lugar de limitar los falsos negativos a nuestro 5% deseado, en realidad los limitamos a 1.27%. Esto dará como resultado el aumento correspondiente en casos veraces como inconclusos. Corregir esto implicará primero calcular el límite alfa corregido usando la forma normal: $1-(1-.05)^4 = .1854$. El uso del alfa corregido de Šidák = .1854 dará lo siguiente: $1-(1-.1854)^{1/4} = .05$. Esto preservará la especificidad de la prueba para veraces con polígrafos exploratorios de asuntos múltiples en niveles aceptablemente altos, al tiempo que reduce la ocurrencia de resultados inconclusos para las personas veraces. También limitará la ocurrencia de errores de prueba falsos negativos a tasas que están dentro del nivel de tolerancia expresados por el nivel alfa = .05.

En la práctica, las correcciones estadísticas se pueden aplicar tanto para el límite alfa o para los valores-p usando las formas normales o inversas. Sin embargo, la corrección de los valores-p solo pueden

lograrse después de conducir y calificar un examen, mientras que la corrección de los límites alfa se puede lograr antes de la realización del examen. Esto se logra al usar la puntuación subtotal con un valor-p igual o inferior al alfa corregido como umbral de decisión.

Conclusión

Todas las pruebas científicas son un proceso de clasificación e inferencia. La clasificación, en este caso, se refiere a la formulación de un resultado de prueba categórico simple. La inferencia es el proceso de calcular una estimación estadística o probabilística respecto de la probabilidad de que hubiera ocurrido un error. En un sentido más abstracto, el propósito de las pruebas científicas es evaluar y cuantificar un fenómeno amorfo que no puede someterse a una observación determinista simple y perfecta o a una medición física / lineal directa. La observación determinista requiere la existencia de ciertos fenómenos que están única y perfectamente asociados con lo que queremos evaluar. Esto sería teóricamente perfecto y también obviaría la necesidad de realizar pruebas. La medición física, en cambio, es casi perfecta, aunque todavía está sujeta a errores mecánicos de medición, y requeriría dos cosas: 1) una sustancia física para medir, y 2) una unidad de medición bien definida. Las pruebas científicas son intrínsecamente probabilísticas: no son deterministas ni son mediciones físicas reales. No se espera que las pruebas científicas sean perfectas. Se espera que cuantifiquen el margen probabilístico de incertidumbre que rodea una conclusión. Las buenas pruebas científicas harán esto de manera que las proporciones previstas del error de la prueba concuerden razonablemente con la evidencia observada de los errores de prueba. Los efectos de la multiplicidad tienen un impacto potencialmente serio en la precisión de las estimaciones del error de prueba. El uso de correcciones estadísticas puede ser una parte importante de la validez y efectividad de un método de prueba.

Dos ideas centrales subyacen a todas las pruebas y experimentos científicos. La primera idea central es que todas las conclusiones científicas o hipótesis son relativas a alguna alternativa. Se espera que los profesionales que hacen conclusiones científicas articulen las alternativas y usen la teoría de la probabilidad para ponderar la evidencia. La segunda idea central es que todas las conclusiones e hipótesis deben expresarse como hipótesis estadísticas o probabilísticas para que puedan ser cuantificables. Las conclusiones o hipótesis que no pueden establecerse como hipótesis estadísticas no pueden medirse o probarse y, por lo tanto, no son científicas.

Se puede decir que las ideas no científicas que pretenden ser científicas son pseudociencia.

En relación con la necesidad de hipótesis verificables, es necesario hacer declaraciones *a priori* sobre la tolerancia al error y el nivel alfa requerido para la significancia estadística. Por lo general, los practicantes de campo no toman decisiones sobre los límites alfa o los cortes de puntuación numéricos - estas son, con frecuencia, políticas de la agencia y se desarrollan en torno a las necesidades específicas del contexto de la gestión de riesgos. Tampoco se espera que los practicantes de campo calculen fórmulas estadísticas. En su lugar, comúnmente usan tablas de referencia estadística publicadas para las que se calcularon previamente todos los posibles resultados de la prueba.

Si la prueba del polígrafo no es más que una herramienta para amplificar o mejorar un interrogatorio o entrevista, los examinadores no necesitarían nunca dar cuenta ni explicar los resultados de la prueba. Si este fuera el caso, ni siquiera necesitarían calificar la prueba, y ciertamente no necesitarían aprender sobre la teoría de la probabilidad y los fenómenos estadísticos. De igual manera, nunca se esperaría que los examinadores poligráficos dieran cuenta o explicaran el resultado de una prueba si se obtuviera una confesión por cada uno de los resultados de prueba de engaño sin excepción. Si la información en las discusiones del pretest y postest son el único propósito de la prueba del polígrafo, entonces no habría necesidad alguna de proporcionar un resultado de la prueba. Sin embargo, si alguna vez hay necesidad de explicar el resultado de una prueba o dar cuenta del nivel de certeza o incertidumbre que debe atribuirse a un resultado de prueba, los examinadores pueden estar obligados a puntuar numéricamente y cuantificar estadísticamente el resultado de la prueba. Los examinadores que no están preparados para hacerlo serán vulnerables a la vergüenza profesional, ya sea por la incapacidad de proporcionar cálculos basados en evidencia de la precisión de la prueba esperada y de la tasa de error, o debido a la frustración cuando finalmente se descubra que los resultados del polígrafo son probabilísticos e imperfectos a pesar de una fingida actitud de certeza.

Los examinadores que están preparados para dar cuenta de los resultados de las pruebas usando los principios básicos, teoría y conceptos de la estadística y la probabilidad estarán mejor preparados para generar impresiones profesionales favorables cuando discutan los resultados de las pruebas sin la sensación de inseguridad que surge de las expectativas ingenuas de la perfección determinista de una prueba. Aunque siempre habrá un valor práctico en la información que se puede obtener de las entrevistas de pretest y postest poligráficas, los resultados de las pruebas sin cálculos realistas de las estimaciones de errores estadísticos, al final, no tendrán ningún valor real.

En última instancia, todas las puntuaciones de prueba, incluidas puntuaciones de gran total y subtotal de las pruebas de polígrafo con preguntas de comparación tendrán una probabilidad de error asociada. La teoría de la probabilidad nos informa que las tasas de error son predeciblemente acumulativas siempre que intentamos hacer comparaciones estadísticas múltiples dentro de una prueba o experimento único. Aunque es preocupante, la predictibilidad de los fenómenos de multiplicidad significa que también podemos aplicar los principios de la teoría de la probabilidad para corregir estadísticamente los efectos de la multiplicidad – si comprendemos los principios de la probabilidad. Si bien los cálculos muy simples, como la corrección de Bonferroni se pueden administrar fácilmente en escenarios de campo, los profesionales de campo deben liberarse de cálculos complejos como la corrección de Šidák mediante la inclusión de información corregida estadísticamente en tablas de referencia normativas publicadas. El uso de algoritmos computarizados también puede lograr la aplicación de estas correcciones estadísticas con confiabilidad automatizada. Aunque muchos investigadores, estadistas y científicos preferirán usar métodos estadísticos

combinados como ANOVA y otros métodos para evaluar simultáneamente hipótesis estadísticas múltiples sin la introducción de los efectos de la multiplicidad, la corrección de Bonferroni y la corrección de Šidák son dos soluciones clásicas para los problemas bien conocidos de la multiplicidad. Son adecuados para el análisis y la interpretación de los resultados de pruebas poligráficas con preguntas de comparación.

Referencias

Abdi, H. (2007). Bonferroni and Sidák corrections for múltiple comparisons. In N.J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics.* Sage.

Department of Defense (2006a). *Federal psychophysiological detection of deception examiner handbook.* Reprinted in Polygraph, 40 (1), 2-66.

Department of Defense (2006b), *Test data anafysis: DoDPI numerical evaluation scoring system.* [Retrieved from http:// www.antipolygraph.org on 3-31-2007],

McDonald, J. H. 2009. *Handbook of Biological Statistics 2nd ed..* Sparky House Publishing, Baltimore, Maryland.

Miller, R. G. (1981). *Simultaneous Statistical Inference 2nd Ed.* Springer Verlag New York.

Nickerson, R. S. (June 1998). "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises". Review of General Psychology 2 (2): 175-220.

Sidák, Z. (1967). Rectangular confidence región for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62,* 626-633.

White, L. A. (August 12, 2011). *Wordofthe Week: Look Elsewhere Effect.* Sanford National Accelerator Laboratory.