

EFFECTO DE UNA VALIDEZ INTERNA IMPERFECTA EN LA INFERENCIA

Vicente Manzano Arrondo (1997)

Resumen

Existen algunos esfuerzos dispersos para obtener medidas de probabilidad en el establecimiento de conclusiones a nivel de los resultados muestrales. Por otro lado, la inferencia estadística suministra estrategias para realizar conclusiones a nivel de la población, conclusiones que vienen acompañadas de medidas de inseguridad en términos probabilísticos. En el presente trabajo, se propone una estrategia formal para compaginar ambos valores de probabilidad y obtener, con ello, una medida global de seguridad o inseguridad para los resultados del estudio.

Palabras clave: Validez, probabilidad, error de primera especie, experimentación.

Introducción

El proceder habitual en la realización de experimentos, así como en su conceptualización teórica, muestra algunas peculiaridades con respecto a los resultados de la inferencia. Como resulta constante dentro de la metodología experimental, tales aspectos se encuentran relacionados con el concepto de validez.

La preocupación principal, constantemente explícita, en la realización de experimentos, se centra en conseguir un control tal que existan garantías suficientes de que los resultados obtenidos son válidos al menos a nivel de la muestra (validez interna). No obstante, el interés científico por establecer leyes generales, exige del experimentador ciertas dosis de preocupación por la aplicabilidad de sus resultados en contextos más amplios (validez externa).

Especialmente desde la publicación del texto de Campbell y Stanley (1966), el binomio validez interna/externa es motivo de discusión casi constante. Es cierto que el concepto de validez ha sufrido cambios y redefiniciones, gozando de mayor aceptación la perspectiva de considerar que los diseños, instrumentos o resultados no son válidos o no por sí mismos, sino que son las interpretaciones de los resultados las que pueden evaluarse como más o menos válidas (Messick, 1989). No obstante, la preocupación por evitar la confusión de variables en el contexto concreto del experimento sigue monopolizando buena parte de la preocupación de los experimentadores.

Es fácil concluir que el grado de validez interna de un experimento tiene efectos sobre los resultados de la inferencia. Si el investigador no está seguro de lo que ha encontrado en el contexto de la muestra ¿Cómo puede medir el nivel de aplicabilidad de los resultados?.

Validez interna y probabilidad

Según el objetivo mencionado de establecer leyes generales, el proceder habitual en ciencia pasa por definir una población, obtener una muestra representativa (aleatoria) de ésta, realizar determinadas operaciones de control, medición y cálculo y poner en marcha un procedimiento de inferencia basado en probabilidades, para terminar estableciendo conclusiones generales con determinados riesgos concretos asociados.

No obstante, es habitual comenzar incorrectamente el proceso, obteniendo muestras no aleatorias, si bien se sigue recurriendo a la inferencia estadística para establecer generalizaciones. Edgington (1995) llama la atención acerca de que obtener muestras no aleatorias es un proceder correcto, considerando que no existen poblaciones concretas en las miras del investigador y que su interés se centra en garantizar la validez al nivel de la muestra (y, como efecto secundario, también al nivel de las poblaciones con respecto a las que la muestra podría considerarse representativa). Por esta razón, las conclusiones se establecen únicamente al nivel del contexto concreto.

Edgington va más allá y suministra algunas reflexiones que quizá no hayan tenido el eco que merecen. En especial, que la inseguridad al nivel de la muestra *puede medirse en términos probabilísticos*. De esta forma pone en marcha las llamadas pruebas de aleatorización, si bien dentro del contexto de los diseños longitudinales, en donde se gestan múltiples alternativas a los procesos de inferencia y probabilidad clásicos (Allison, Franklin y Heshka, 1992).

El aspecto que más nos interesa de esta discusión surge de considerar simultáneamente las perspectivas de Campbell y Stanley, Messick y Edgington: *la validez interna, tomada como una cuestión de grado, puede ser medida en términos probabilísticos*. Si ello fuera posible en cualquier diseño específico, el investigador podría

suministrar una medida numérica concreta de cuán seguro se encuentra de las conclusiones al nivel de la muestra.

Una medida global de probabilidad para la validez de un experimento

Asociar la validez externa y la probabilidad es el más antiguo y principal logro de la inferencia estadística: permite generalizar los resultados desde la muestra a la población, acompañando una medida (probabilidad) de seguridad o inseguridad en las conclusiones. Pero en la base de este procedimiento general se encuentra la suposición de que los resultados obtenidos en la muestra son incuestionables a nivel de ésta. En otros términos: existe un grado máximo de validez interna.

En multitud de situaciones, especialmente en la aplicación de diseños cuasiexperimentales, existe cierto nivel de inseguridad incluso en la elaboración de las conclusiones a nivel de la muestra. Tal circunstancia debería ser considerada a la hora de realizar una inferencia con los resultados obtenidos.

Para formalizar aritméticamente una medida global de inseguridad, utilizaremos los símbolos:

- p(m) probabilidad de errar al afirmar la existencia de efecto o relación, en el contexto de la muestra
- p(i) probabilidad de errar al afirmar la existencia de efecto o relación, en el momento concreto de la inferencia estadística.
- p(t) probabilidad de errar al afirmar la existencia de efecto o relación, en el proceso global de la investigación.

Luego, la probabilidad global de errar al afirmar la existencia de efecto o relación, considerando ambos procesos, será:

$$p(t) = 1 - ([1 - p(m)][1 - p(i)])$$

En el caso de mantener la suposición de una validez interna de grado máximo, $p(m)=0$, ocurrirá $p(t)=p(i)$, con lo que la probabilidad global se encontrará localizada únicamente en el proceso de inferencia en sí.

Existen dos alternativas para aplicar esta estrategia en el momento de tomar una decisión, dentro de una prueba de significación.

1. Modificar el valor obtenido para el grado de significación ($p(O/H_0)$ en Manzano, 1997), realizando la comparación habitual con el nivel de significación (α). En tal caso:

$$p(O/H_0) = 1 - ([1 - p(m)][1 - p(i)])$$

2. Modificar el referente de comparación α , manteniendo el habitual $p(O/H_0)=p(i)$ mediante

$$\alpha' = 1 - ([1 - p(m)][1 - \alpha])$$

$$\alpha' = 1 - \frac{1 - \alpha}{1 - p(m)}$$

donde α representa el valor de comparación en el momento de la inferencia y α' el valor global.

Como puede intuirse, la inseguridad global no puede ser inferior a la que se calcule para cualquiera de las etapas constituyentes del proceso; es decir, no puede establecerse un valor para α' que sea inferior a $p(m)$. Dado que $\alpha \geq 0$, entonces

$$\frac{1 - \alpha'}{1 - p(m)} \geq 1 \quad \text{Y} \quad \alpha' \geq p(m)$$

Un ejemplo concreto

Manzano, Pérez Santamaría y Fazeli (1997) pusieron a prueba un procedimiento de enseñanza por ordenador, mediante un cañón de imágenes. Para esta experiencia se utilizaron cinco grupos de clase, subdivididos cada uno de ellos en dos subgrupos de prácticas. Dentro de cada grupo de clase se decidió aleatoriamente cuál de los subgrupos recibiría el tratamiento (clase mediante el cañón de imágenes) y cuál sería el control (clase mediante pizarra). Aplicada una prueba de significación apropiada, se concluyó que existían diferencias entre ambos procedimientos, generándose un mayor aprendizaje en los que recibieron la clase mediante cañón de imágenes. El grado de significación obtenido fue $p(O/H_0) = 0.0015$, y $\alpha = 0.05$ el nivel de significación utilizado.

La clase versó sobre la teoría de la decisión estadística. Era la primera sesión para tal tema y los alumnos de la muestra no habían tomado contacto con esta materia con anterioridad (los repetidores no fueron considerados en los análisis). Con propiedad no fue un experimento, sino un *cuasiexperimento*, pues se trató con grupos naturales. La validez interna no es impecable. Se aleja del grado máximo en una distancia inicialmente desconocida. Sin embargo, alguna luz puede arrojar al respecto.

Fueron grupos naturales, ello implica que aparentemente puede existir un nivel diferente de los grupos en lo que se refiere a la variable dependiente. Podría haberse aplicado un pretest, no obstante resultaba innecesario, ya que se suponía desconocimiento absoluto sobre el tema en todos los componentes de la muestra. Otro factor múltiple pudo haber influido diferencialmente, generando confusión de variables: algunos de los grupos pudieron aprehender los conceptos con mayor aprovechamiento que otros, no por el procedimiento utilizado sino por sus características propias, como pudo ser una capacidad

diferencial de aprendizaje.

Si la experiencia se hubiera realizado sobre uno solo de los grupos de clase, no podría asegurarse si las diferencias observadas entre los dos subgrupos se debían al tratamiento o a las características de cada conjunto de sujetos. Ante un desconocimiento total sobre ello, podría asignarse una probabilidad de 1/2 para cada una de las dos alternativas de explicación. Dado que se manejaron cinco grupos, puede establecerse

$$p(m) = 1/2^5 = 0.03125$$

Así pues, con una probabilidad de valor $p(m)=0.03125$, pudo ocurrir que los resultados obtenidos se debieran a las características de los grupos y no al efecto del tratamiento. ¿Qué ocurre ahora con la generalización o aplicabilidad de los resultados?. Si bien el proceso de inferencia no tiene por qué variar, deben corregirse las probabilidades implicadas con el objetivo de que el procedimiento total considere que la etapa muestral no se ha cubierto con una validez interna de grado máximo. Al respecto, pueden utilizarse cualquiera de las dos estrategias que se han hecho explícitas en un apartado anterior:

1. Modificación del valor para el grado de significación antes de tomar una decisión sobre el rechazo o mantenimiento de la hipótesis de nulidad:

$$p(O/H_0)' = 1 - (1 - 0.0015)(1 - 0.03125) = 0.0327$$

Dado que $p(O/H_0)=0.0327 < 0.05=\alpha$, se procede al rechazo de H_0

2. Modificación del valor para el nivel de significación. El nuevo valor para el acto aislado de la decisión estadística, será decidido en función del valor que se decide para el nivel máximo de inseguridad para el conjunto del proceso. Si éste se mantiene en $\alpha' = 0.05$, entonces:

$$\alpha' = 1 - \frac{1 - 0.05}{1 - 0.03125} = 0.0194$$

Dado que $p(O/H_0)=0.015 < 0.0194=\alpha$, se procede al rechazo de H_0

Discusión

El sentido original del proceso de inferencia es el de contar con una medida de seguridad acerca de la estimación o decisión que toma el investigador con respecto a unos resultados concretos. Tradicionalmente, la puesta en práctica de este procedimiento ignora el nivel de seguridad de los resultados en la muestra. Razonablemente, cuanto mayor sea la creencia del investigador en que no se han controlado influencias diferenciales en el estudio, mayor debería ser la inseguridad en el último estadio, el de traslación de los resultados hasta la población, real o hipotética. El objetivo de este trabajo ha sido el de aportar algunas herramientas útiles para trabajar en la dirección mencionada.

Sin embargo, las propuestas que aquí se presentan son insuficientes. Las casuísticas son múltiples y la obtención de un nivel de probabilidad que exprese el grado de validez interna es harto dificultoso. No obstante, la existencia de problemas fácticos no debe restar valor a la creencia de que tales inconvenientes son temporales y que un esfuerzo de tenacidad e imaginación metodológicas debe generar nuevas soluciones para situaciones concretas, antesala para las soluciones generales.

Lo importante, en definitiva, es que contamos con los recursos suficientes para manejar una medida global de inseguridad $p(t)$, si el investigador es capaz de tomar una decisión sobre α y obtener un valor concreto para $p(m)$.

Referencias

- Allison, D.B.; Franklin, R.D. y Heshka, S. (1992). Reflections on Visual Inspection, Response Guided Experimentation, an Type I Error Rate in Single-Case Designs. *Journal of Experimental Education*, vol 61, num 1, págs 45-51.
- Campbell D.T. y Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNall & Company.
- Edgington, E.S. (1995) *Randomization Tests*. Third Edition. xxxxx
- Manzano Arrondo, V. (1997). Usos y abusos del error tipo I. *Psicológica*, xxxxx.
- Manzano Arrondo, V.; Pérez Santamaría, F.J. y Fazeli Khalili, H. (1997). Aprendizaje por ordenador: Una aplicación para la enseñanza de la teoría de la decisión estadística. *Actas del V Congreso de Metodología de las Ciencias Humanas y Sociales*. Sevilla (en prensa).