Contents

Resumen

Luego de años de desarrollo y ejecución, después de sortear obstáculos prácticos, teóricos y de inferencia de la investigación experimental en las ciencias sociales, su experimento ha finalizado. Al comparar los grupos de tratamiento y de control, encuentra un resultado sustancial y estadísticamente significativo en una variable de resultado de interés teórico. Antes de que pueda abrir la champaña para celebrar, un amable colega le pregunta: "Pero ¿qué nos dice esto sobre el mundo?".

1. ¿Qué es la validez externa?

La validez externa es un nombre que se da a la generalización de resultados, y consiste en preguntarse "si una relación causal se mantiene a pesar de la variación de personas, entornos, tratamientos o variables de resultado"[1]. Por ejemplo, una de las preocupaciones frente a la validez externa es si los experimentos de laboratorio tradicionales en economía o psicología llevados a cabo con estudiantes universitarios, producen resultados que son generalizables al público común. Por ejemplo, en la economía política del desarrollo, podríamos evaluar la posibilidad de que un programa de desarrollo impulsado por la comunidad en la India, pueda aplicarse (o no) en África Occidental o en América Central.

La validez externa adquiere especial importancia cuando se formulan recomendaciones sobre políticas públicas derivadas de la investigación. La extrapolación de los efectos causales de uno o más estudios a un contexto político determinado requiere una cuidadosa consideración tanto de la teoría como de las evidencias empíricas. En esta guía de métodos se analizan algunos conceptos clave, los obstáculos que hay que evitar y las referencias útiles que hay que tener en cuenta al pasar de un Efecto Promedio Local a un mundo más amplio y tangible.

2. ¿En qué se diferencia de la validez interna?

La validez interna se refiere a la calidad de las inferencias causales que se hacen para un determinado grupo de sujetos. Tal y como planteó Campbell[2], la validez interna pregunta: "¿tuvo el estímulo experimental

[1] Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company.

[2] Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological bulletin, 54(4), 297.

algún impacto significativo en este caso específico?". Este concepto encaja con el enfoque contrafactual de la causalidad que suelen utilizar los experimentalistas, el cual se pregunta si las variables de resultado cambian en función de la presencia o ausencia de un tratamiento[3].

Antes de poder extrapolar un efecto causal a una población distinta, es vital que el Efecto Promedio del Tratamiento original se base en un resultado bien identificado. Para la mayoría de los experimentalistas, la asignación aleatoria proporciona la variación de identificación necesaria, siempre que no haya desgaste, interferencia, efectos secundarios u otras amenazas a la inferencia. En el caso de los estudios observacionales, se necesitan supuestos de identificación adicionales, como la independencia condicional del tratamiento de las variables de resultado potenciales.

3. Cómo compensar entre la validez interna y externa

Existe un debate en el campo de las ciencias sociales sobre la importancia de identificar resultados internamente válidos (que por definición se aplican a una muestra local), y de generar resultados que puedan extrapolarse a poblaciones de interés más extensas. Resulta útil familiarizarse con este debate al considerar las decisiones que inevitablemente hay que tomar en materia de diseño cuando se trata de intervenciones con recursos limitados. El hecho de que en ambos lados del debate se incluyan expertos en econometría da cuenta de la importancia de este tema.

En un lado de la discusión se encuentran los defensores de "primero la identificación", quienes argumentan que, sin resultados internamente válidos, un estudio simplemente no aporta información útil, independientemente de si se trata de una población o un contexto local o general. Como dice Imbens[4], "sin una fuerte validez interna, los estudios tienen poco que aportar a los debates de política pública, mientras que los estudios [internamente válidos] con una validez externa muy limitada a menudo son (y en mi opinión deberían ser) tomados en serio en dichos debates."

Del otro lado del debate, expertos sostienen que incluso sin la plena identificación de un resultado internamente válido, se puede rescatar información útil, especialmente si es relevante para cuestiones importantes que afectan a un contexto amplio. Manski[5] escribe que "lo que importa es la informatividad de un estudio para la elaboración de políticas públicas, que depende conjuntamente de la validez interna y externa". Manski asegura que con los datos de un estudio amplio pero pobremente identificado, se pueden generar límites sobre la estimación de interés que, aunque no son tan útiles como una estimación puntual precisa, siguen haciendo avanzar a la ciencia.

4. Teoría y generalización

Extrapolar un resultado a un contexto, a una variable de resultado, a una población o a un tratamiento distintos no es un proceso mecánico. Tal y como afirman Samii[6] y Rosenbaum,[7] la teoría pertinente debe utilizarse para guiar la generalización, tomando las pruebas existentes relevantes y haciendo predicciones para otros contextos de forma fundamentada. Las teorías reducen los problemas complejos a representaciones más suaves y ayudan a dilucidar qué factores son importantes. Del mismo modo que la teoría orienta el contenido de las intervenciones y los diseños de investigación, también puede decirnos qué elementos del alcance de una teoría son relevantes para extrapolar un resultado. ¿Qué covariables importan? ¿Qué información contextual importa?

[3]Se pueden encontrar más detalles en la guía de métodos de inferencia causal: http://egap.org/resource/10-cosas-que-debe-saber sobre-la-inferencia-causal

[4]Imbens, G. (2013). Artículo de reseña del libro: Public Policy in an Uncertain World: By Charles F. Manski. The Economic Journal,123(570), F401-F411.

[5]Manski, C. F. (2013). Response to the review of 'public policy in an uncertain world'. The Economic Journal 123: F412-F415.

[6]Samii, Cyrus. (2016). "Causal Empiricism in Quantitative Research." Journal of Politics 78(3):941–955.

[7]Rosenbaum, Paul R. (1999). "Choice as an Alternative to Control in Observational Studies" (con discusión). Statistical Science 14(3): 259–304.

5. ¿Cómo puedo determinar dónde aplican mis resultados?

Existen dos formas principales para generalizar resultados; una basada en las covariables de las unidades del estudio, y otra basada en la manipulación experimental real de las variables moderadoras. Al observar cómo varía el efecto del tratamiento con respecto a una variable no aleatoria pre-tratamiento puede describir la heterogeneidad del efecto del tratamiento, lo cual puede sugerir para quién o en dónde la intervención sería más eficaz, más allá de la muestra original. Sin embargo, hay que tener en cuenta que este tipo de análisis no puede determinar si la heterogeneidad del efecto del tratamiento es causada por esa variable pre-tratamiento. La preocupación -propia de la investigación observacional- es que la covariable no aleatoria puede estar correlacionada con una variable no observada, y es este factor "no visto" el que de hecho es responsable de los impactos heterogéneos del tratamiento[8]. Lo ideal es aprovechar la variación exógena en el moderador de interés, descartando así la posibilidad de tal confusión. Un diseño experimental factorial en el que el investigador asigna el moderador independientemente del tratamiento principal de interés, puede generar pruebas especialmente convincentes sobre el papel de un moderador. Aunque, por supuesto, las consideraciones de coste y el poder estadístico pueden impedir este enfoque en la práctica.

Dado que la generalización es principalmente un ejercicio de predicción, donde se pregunta en qué lugar podemos esperar una relación causal similar a la observada localmente, la extrapolación de efectos heterogéneos basada en covariables similares suele ser razonable, siempre y cuando la teoría no indique fuentes de confusión[9]. No obstante, la prueba más sólida de la posibilidad de generalización de un resultado proviene de una interacción bien identificada entre un moderador exógeno y el tratamiento, propagada después a través del perfil de covariables de una población objetivo. De hecho, con algunas hipótesis sólidas, la extrapolación puede proporcionar resultados tan buenos o mejores que la realización de un segundo experimento in situ[10] El cálculo de una estimación extrapolada a menudo puede realizarse mejor mediante "machine learning", aunque la regresión lineal también funciona razonablemente bien[11].

6. El comportamiento estratégico puede sabotear sus extrapolaciones

La extrapolación de un resultado local a un contexto diferente puede resultar difícil, incluso con un perfil de covariables convincente al que se quieran generalizar efectos. Una manipulación experimental aleatoria a nivel local genera un "efecto de equilibrio parcial". Las dinámicas estratégicas (que incluyen los reveses o comportamientos compensatorios) fuera del contexto local de una intervención experimental, pueden complicar los esfuerzos para generalizar un resultado. Supongamos, por ejemplo, que se demuestra que una intervención de transferencia en efectivo incondicional aumenta el bienestar, el espíritu empresarial y el empleo en una muestra de 200 pueblos. ¿Qué pasaría si la intervención se ampliara para abarcar 1.000 pueblos? En este punto, podríamos imaginar que las regiones excluidas del programa tienen más probabilidades de escuchar sobre él. Las unidades no tratadas podrían empezar a pedir del gobierno otro tipo de transferencias, dando lugar a efectos similares a los producidos por la transferencia directa de efectivo. De forma similar, a veces las relaciones causales sólo funcionan cuando se aplican en ciertas personas. Por ejemplo, imaginemos un programa de capacitación laboral que funciona muy bien (en comparación con los trabajadores que no recibieron el programa), ¿qué pasaría si se extendiera a todos los trabajadores? Incluso si hay efectos positivos en todos los participantes, podría haber efectos reducidos o nulos en promedio, ya que los puestos de trabajo más calificados ya están ocupados por la primera tanda de trabajadores, y la segunda tanda se vería obligada a permanecer en sus empleos anteriores, aunque ahora de manera sobrecalificada. En resumen, en condiciones de equilibrio general podríamos esperar resultados diferentes incluso cuando el perfil de las covariables coincide.

[8]Gerber, A. S., & Green, D. P. (2012). Field experiments: Design, analysis, and interpretation. WW Norton.

[9]Bisbee, James; Rajeev Dehejia; Cristian Pop-Eleches & Cyrus Samii. (2016). "Local Instruments, Global Extrapolation: External Validity of the Labor Supply-Fertility Local Average Treatment Effect." Journal of Labor Economics

[10]Bisbee, James; Rajeev Dehejia; Cristian Pop-Eleches & Cyrus Samii. (2016). "Local Instruments, Global Extrapolation: External Validity of the Labor Supply-Fertility Local Average Treatment Effect." Journal of Labor Economics

[11]Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. Journal of Research on Educational Effectiveness, 9(1), 103-127.

7. No hay que confundir la validez externa con la validez de constructo o la validez ecológica.

La validez interna y externa no son los únicos problemas de "validez" que pueden plantearse en el trabajo experimental y, aunque son relevantes, también son distintos. La validez ecológica, tal y como la definen Shadish, Cook y Campbell[12] se refiere a si una intervención parece artificial o fuera de lugar cuando se sitúa en un contexto nuevo. Por ejemplo, ¿un taller de información llevado a cabo por experimentadores en un pueblo rural, se podría parecer a los tipos de intercambio de información que la población puede experimentar en la vida normal? Del mismo modo, si el mismo taller se realizara en una ciudad grande, ¿parecería fuera de lugar?

La validez de constructo considera si un concepto teórico puesto a prueba en un estudio se operativiza adecuadamente mediante el tratamiento o los tratamientos. Si su experimento pone a prueba el efecto de la ira en la reciprocidad política, y en realidad está manipulando el miedo o la confianza en su tratamiento, la validez de constructo puede ser afectada. Tanto la validez de constructo como la ecológica son relevantes para las generalizaciones y, por lo tanto, son útiles para hacer afirmaciones sobre la validez externa.

8. Extrapolación entre tratamientos y variables de resultado

Aunque gran parte de esta guía se ha centrado en la portabilidad de un determinado tratamiento a un nuevo lugar o momento, la validez externa también considera las variaciones en los tratamientos y las variables de resultado. Es decir, imaginemos que hacemos el mismo experimento sobre la misma muestra, pero con una variación en el tratamiento, ¿podríamos predecir que el efecto causal local será similar? Del mismo modo, ¿podemos predecir si un determinado tratamiento producirá los mismos o diferentes efectos causales en una variable de resultado distinta? A veces podemos abordar estas preocupaciones al realizar experimentos que evalúan tratamientos y variables de resultado alternativos. Cuando los experimentos de seguimiento son escasos, estas cuestiones deben resolverse analíticamente. En este caso la extrapolación requiere pensar, con ayuda de la teoría, en las características de los tratamientos o las variables de resultado y hacer predicciones razonables.

9. La replicación es importante

Ningún estudio por sí solo representa la "última palabra" sobre una pregunta académica. Siguiendo la lógica de la actualización Bayesiana, las evidencias adicionales a favor o en contra de una teoría determinada permiten a la comunidad científica y política actualizar sus opiniones sobre la fuerza y la validez de una relación causal.

La replicación de los estudios es una parte importante de este proceso de actualización: los académicos deben replicar los estudios en contextos que parecen muy diferentes, pero también en algunos contextos que parecen muy similares. Lo primero nos permite identificar relaciones causales locales que pueden triangularse con evidencias existentes y generalizarse según convenga. Al mismo tiempo es importante replicar directamente los estudios existentes en condiciones lo más parecidas posible al original, para verificar que los efectos locales que uno puede estar interesado en extrapolar son realmente fiables. Por ejemplo, La Open Science Collaboration[13] descubrió que al reproducir 100 importantes experimentos de psicología, sólo el 47% de los tamaños de los efectos comunicados originalmente estaban dentro del 95% del intervalo de confianza del tamaño del efecto mostrado en la réplica.

[12]Shadish, W. R., Cook, T. D., y Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company.

[13]Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

10. Tenga en cuenta el tiempo

Al pensar en las relaciones causales de interés, es importante también tener en cuenta el tiempo: ¿las cosas que aprendemos sobre el pasado se extienden al futuro? ¿Cómo cambian los resultados potenciales para un individuo con el tiempo? Leyes inmutables gobiernan el mundo físico y químico; por tanto lo que aprendemos hoy sobre estas leyes siempre será verdadero. En cambio, entendemos mucho menos sobre las motivaciones subyacentes al comportamiento social y si se mantienen constantes durante el tiempo. La respuesta puede ser negativa. A la hora de tomar decisiones sobre la relevancia de políticas públicas y la generalización de los resultados, estas consideraciones pueden ayudar a los académicos a determinar un nivel razonable de incertidumbre y permitir a los responsables de las políticas públicas a ajustarse en consecuencia.