REVISIÓN DEL LIIBRO "BEYOND SIGNIFICANCE

TESTING, STATISTICS REFORM IN THE BEHAVIORAL SCIENCES" DE REX KLINE (SEGUNDA EDICIÓN, 2013, WASHINGTON, AMERICAN PSYCHOLOGICAL ASSOCIATION, ISBN: 978-1-4338-1278-1)

Claudia Patricia Ovalle-Ramirez / cpovaller@unal.edu.co

Egresada de Psicología
Universidad Nacional de Colombia
Investigadora
Pontificia Universidad Católica de Chile

Rex Kline, de la Universidad de Concordia, Montreal, argumenta en su libro: Beyond Significance testing, statistics reform in the behavioral sciences, que la forma en la que evaluamos hipótesis está afectando el desarrollo de la Psicología y otras ciencias empíricas. Pero ¿por qué? Porque resultados estadísticamente significativos no lo son siempre en la práctica (la significancia clínica, la relevancia científica o el tamaño del efecto puede ser 0). De hecho, Kline afirma que, como ciencia empírica, la Psicología puede estar convirtiéndose en una ciencia de los "falsos positivos"; esto, a pesar del repetitivo consejo de la American Psychological Association (APA), de que cada estudio debe proveer evidencia de que tiene el poder para detectar efectos de interés substantivo (APA, 2001).

Según la literatura, los estudios no experimentales, en promedio, tienen un 50 % de poder explicativo, dejando el 50 % al azar. Lo anterior ocurre por dos motivos sustanciales, de acuerdo con el argumento de Kline: primero, la crisis de medición que implica la ausencia de reporte de la confiabilidad de las medidas o el reporte inductivo de estas medidas (es decir, se deduce de manuales u otras fuentes indirectas). Segundo, debido a que los investigadores no se están asegurando de que sí se respetan los supuestos de las técnicas estadísticas que se emplean (ej. los relacionados con la distribución de los datos).

Otros fallos comunes de los investigadores son, según Kline: la falta de reporte sobre valores perdidos, como se resuelve este hecho, y la falta de reporte de las decisiones que toman (sobre métodos, datos, fuentes y alternativas de análisis disponibles). Para el caso de los psicólogos, es importante tener en cuenta que la insuficiencia de los "p" valores también se debe a que la mayoría de los estudios no son basados en muestras aleatorias y que se asumen fuentes de variación (error) que no se hacen explícitas.

Por lo que, en la misma línea, Kline argumenta que los valores p no son suficientes por su imprecisión. Por ejemplo, el siguiente output de SPSS tomado de Kline (p. 21) es una muestra de que los valores p no son valores precisos necesariamente:

$$t\,(27) = 2{,}73, p = 0.250001840178221007$$

En este ejemplo, dado que "el valor p no describe hipótesis sino datos" (Kline, 2013, p. 14), Cumming (2012, en Kline 2013) habla de la nueva estadística, la cual está compuesta por el uso de tamaños del efecto e intervalos de confianza para reemplazar el sobre uso (y el abuso) de los valores p. Esto con la excepción de los casos en los que la variable 'respuesta' está dada en términos de rango o en una estructura jerárquica compleja. La importancia del intervalo se encuentra en que este reporta con exactitud la medida del error en el cálculo muestral del estimador estadístico. En consecuencia, indica Kline, incluir el estimador con su intervalo es una forma de desarrollar pensamiento estimativo [estimation thinking] en donde el investigador se acostumbra, no solo a reportar cuánto (el estadístico), sino que también qué tan preciso es el estimador (el margen de error), además de la significancia práctica, teórica o clínica de los hallazgos.

Sin embargo, puede ocurrir que, tanto el estimador, como el intervalo sean erróneos y no se supere las dificultades que se encuentran cuando se reporta solo p. En este caso, dice Kline, no hay más solución que un juicio experto sobre la significancia substantiva de los hallazgos. Antes del uso de los p valores, se hacía empleo de la tasa crítica del estadístico muestral sobre su error estándar, lo que ahora se llama puntaje z o t (cuando hay normalidad); así mismo, en campos como las Ciencias, se empleaba el test estadístico para detectar outliers, mas no para probar hipótesis. A la aproximación actual de la Psicología se le conoce como "t-estimation" (responder a la pregunta ¿ ?), lo cual puede conllevar a error puesto que un p de .025 no necesariamente significa que la hipótesis nula sea cierta un 2,5 % de las veces.

Continuamente, en los artículos científicos, el lenguaje de "t-estimation" se reproduce con una gran cantidad de números y efectos significativos, los cuales no se inter-

pretan, y, comúnmente, los tamaños del efecto no se reportan; esto se conoce como "sizeless science" (Kline, 2013). Con ello presente, desde 2001, en la quinta edición del Manual de Estilo, la APA generó tres recomendaciones relacionadas con el exceso de artículos de investigación que tienen un perfil de "sizeless science", y que aún se mantienen:

i. Reportar estadísticos descriptivos como medias de varianzas y tamaños de cada grupo, y reportar la matriz de varianza-covarianza intragrupos en los estudios comparativos.
ii. Los tamaños del efecto deben reportarse (a excepción de casos como variables ordinales).
iii. El uso de intervalos de confianza se recomienda pero no es requerido.

Las revistas que requieren tamaños del efecto se encuentran Health Psychology, Journal of Educational Psychology, Journal of Experimental Psychology: Applied. Para Kline, la buena práctica consiste en disminuir el rol de los tests de significancia, desarrollar la replicación como método estándar, y movernos de la confirmación de hipótesis a la prueba de modelos. A continuación se presenta el desarrollo que hace Kline sobre el tema de error en el muestreo y del uso de los intervalos de confianza para evitar la dependencia en el valor p.

Según Kline, los tres mensajes claves sobre muestreo son:

i. El error muestral afecta a todos los estadísticos.
ii. La estimación por intervalos aproxima los márgenes de error.
iii. Hay otras fuentes de varianza o error que no deben omitirse.

Es importante que, en Psicología, se tenga en cuenta que el muestreo aleatorio en cualquiera de sus formas no es garantía de que la muestra sea representativa de la población. Según Kline, el modelo de inferencia a la población se sostiene si las muestras se obtienen de forma aleatoria y por medio de replicaciones (validez externa). De acuerdo con la ley de los grandes números [large numbers], las muestras más grandes tenderán a producir estadísticos más precisos del parámetro de la población. En la estimación de parámetros, las medias muestrales son estimadores, los cuales reducen el sesgo ya que su valor es aproximadamente del mismo tamaño al de la media de la población, al igual que la suma de cuadrados.

$$\sum(X - M) = 0 \text{ y la } SS = \sum(X - M)^2$$

Sin embargo, otros estimadores, como la varianza muestral, derivada como $S^2 = \frac{SS}{N}$ en lugar de la estimación muestral $S^2 = \frac{SS}{df}$, es un estimador sesgado negativamente ya que tiende

a ser menor que el valor poblacional σ^2 Por su parte, el error muestral varia de acuerdo con el tamaño muestral, generando errores mayores para muestras más pequeñas, como se observa en la fórmula $S_M = s/\sqrt{N}$.

A estas fuentes de error explicitadas por Kline, se añaden otras fuentes como los errores de medición, los errores en la definición del constructo (constructos que no miden lo que creen medir), errores de especificación (omisión de variables importantes) y errores de implementación en el tratamiento.

Intervalo de confianza

El intervalo se construye por medio de la adición y resta a un estadístico del producto de su error estándar y el valor critico al nivel de significancia. Este producto es el margen de error. La fórmula del intervalo de confianza para la media es:

$$M \pm S_M [t_{2colas, \alpha} (N - 1)]$$

Los intervalos se interpretan considerando que pueden no contener el valor de la media de la población, dado que este intervalo es susceptible al error muestral. De este modo, no es correcto considerar que un intervalo contiene la media con una probabilidad de 95 %, ya que cada intervalo tiene una probabilidad de 0 % o 100 % de contener esa media (esté contenida o no). Asimismo, el intervalo debe ser interpretado de forma cautelosa, porque según Kline un intervalo puede significar una gran diferencia en términos prácticos (p.ej., si se considera en términos de una dosis de un medicamento letal).

Posteriormente, Klein presenta la estimación de un intervalo de confianza para la diferencia entre las medias de dos poblaciones independientes μ_1-μ_2.

En el ejemplo, un diseño de 10 casos por grupo con $M_1 = M_1 = 13, S_1 = S_1 = 7{,}50$ y $M_2 = M_2 = 11, S_2 = S_2 = 5{,}0$. Donde el intervalo de confianza está dado por:

$$(M_1 - M_2) \pm SM_1 - M_2 [t_{2colar, \alpha} (N - 2)]$$

Donde $SM_1 = M_2$ es el error estándar de la diferencia de medias, calculado como la raíz cuadrada del promedio ponderado de las varianzas para cada grupo, o:

$$S^2 pooled = \sqrt{\frac{2\left(\frac{(S_1^2) + S_2^2}{2}\right)}{n}}$$

Con esto, el intervalo de confianza del 95 % obtenido es $2 \pm 1{,}118\ (2{,}101)$. Este equivale a $[-0{,}35{,}435]$ que incluye el número 0, pero puede o no incluir el valor respectivo de la diferencia de las medias en la población. Usualmente, los investigadores asumen la inclusión del 0 en el intervalo como no significativo. Para ello, existe una regla descrita por Kline que indica que si existe un cotejo entre el intervalo de confianza de la estimación de la media de dos grupos, entonces no hay diferencia significativa al nivel correspondiente de alfa.

Klein sugiere que se considere que los intervalos de confianza no se pueden confundir con test estadísticos, ya

que, entre otras razones, las hipótesis nulas son requeridas para los test estadísticos y no para los intervalos; igualmente, se ha encontrado que existen hipótesis nulas implausibles, las cuales se han aportado en las ciencias (p.ej., equivalencia en la probabilidad de mortalidad en especies jóvenes y adultas). En cambio, los intervalos de confianza ofrecen, si se les replica varias veces, una estimación menos susceptible de sesgo, comparado con los test estadísticos. Por ello, uno de los aportes de Kline es sugerir que se haga un proceso de replicación de intervalos de confianza, a fin de sustituir la práctica común de los psicólogos, de tomar el valor p del primer resultado obtenido.

Un ejemplo propuesto por Kline, consiste en presentar seis replicaciones hipotéticas de comparaciones entre dos muestras, cada una de 20 individuos, con alfa de 0,5 y de las cuales se obtiene la diferencia de medias M1 y M2. Cada una de las pruebas de hipótesis arroja un resultado diferente (tres rechazos de la hipótesis nula y tres pruebas que aceptan dicha hipótesis), con diferencias de medias que varían desde 2,50 hasta 5,0. En este caso, se propone hacer un promedio de la diferencia de medias con todos los datos M_1 - M_2, de modo que, al emplear todos los datos, se obtenga una estimación de la diferencia de medias más adecuada (el autor lo llama una media metanalítica *[meta analytic weighted average]).* En este ejemplo, la media es 3,54 con un intervalo de confianza del 95 % entre (.53, 4.54). Esta media es mejor, dado que la cantidad información contribuye al intervalo de confianza, con base en los resultados promediados de todas las muestras (replicaciones).

En dicho caso, el procedimiento de Welsch se puede emplear en el cálculo de intervalos de confianza, cuando no hay homocedasticiad. Este procedimiento arroja intervalos menos anchos y puede no ser funcional cuando la distribución de los datos no es normal o cuando los grupos son menores a 30 datos. El procedimiento para el cálculo del error estándar de una diferencia de medias es:

$$\text{Swel} = \sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}$$

Esta diferencia, calculada con una varianza agrupada [pooled within group variance], enmascara las diferencias entre los dos grupos (no permite diferencias grandes entre las varianzas), pero el procedimiento de welch permite que los grupos tengan varianzas diferentes. En consecuencia, los grados de libertad para el estadístico t se calculan como:

$$df\text{Wel} = \frac{\left(\frac{S_1^2}{n1} + \frac{S_2^2}{n2}\right)^2}{\left(\frac{(S_1^2)^2}{n_1^2(n_1 - 1)} + \frac{(S_2^2)^2}{n_2^2(n2 - 1)}\right)^2}$$

La forma general de un intervalo de confianza 100(1-) % para la diferencia entre μ_1-μ_2. es la siguiente:

$$M_1 - M_2 \pm SWel\ [\ t_{2\text{-tail},}\ \alpha\ (df Wel)]$$

Si el valor de t obtenido no es un entero, entonces, funciones como TINV de Excel o las calculadoras web pueden indicar el valor crítico de t dados un nivel alfa y los *df* .(Kline, 2013)

Conclusión

El libro concluye que la nueva estadística y la Psicología, como ciencia, no pueden seguir dependiendo del reporte de los p valores, debido a que los resultados estadísticamente significativos no lo son siempre en la práctica, lo cual puede llevar a falsos positivos. Es necesario que se implementen nuevas prácticas en la investigación y en el reporte de resultados, como el uso de intervalos de confianza y estimaciones del tamaño del efecto. Estas prácticas deben extenderse necesariamente a la formación de los psicólogos colombianos y se deben ver reflejadas en las tesis de pregrado y postgrado.

Referencias

American Psychological Association (APA). (2001).Manual de estilo de publicaciones de la Asociación Americana de Psicología (quinta edición). Washington D.C.: APA.

Kline, R. (2013). *Beyond significance testing. Statistics reform in the behavioral sciences* (segunda edición). Washington, D.C.: American Psychological Association.