EVALUAR

Laboratorio de Evaluación Psicológica y Educativa

Evaluar, 14 (2014), 43 – 66
ISSN 1667-4545

Estimación por intervalo del tamaño del efecto expresado como proporción de varianza explicada

Eduardo Bologna*

* Facultad de Psicología, Universidad Nacional de Córdoba, Argentina

Resumen.

En la actualidad se observa un uso más intenso de indicadores de tamaño del efecto para informar resultados de investigaciones. Se discute primero la importancia de comunicar de manera comparable los resultados de investigación en Ciencias Sociales, para luego ejemplificar diferentes casos de aplicación de medidas de tamaño del efecto. A continuación se ilustra la razón por la que la estimación del tamaño del efecto no es una extensión intuitiva de los intervalos de confianza usuales y se muestra el procedimiento para hacerla. Finalmente, se usa el paquete MBESS en lenguaje R a través de INFOSTAT® para realizar las estimaciones de manera sencilla.
Palabras clave: Tamaño del Efecto – Intervalos de Confianza – Software Estadístico

Abstract:
In recent years there has been an increasing use of effect size measures to report results. The current article addresses various issues: first, the importance of compare results in communicating social research is discussed. Second, examples of applications of different measures of effect size are provided. Third, reason supporting that the estimation of effect size is not simply extension of usual confidence intervals are illustrated. Lastly, MBSESS R language by INFOSTAT® for simply estimate.
Keywords: Effect Size - Confidence Intervals - Statistical Software

Introducción

En la actualidad se observa un uso más intenso de indicadores de tamaño del efecto para informar resultados de investigaciones (García García, Ortega Campos & De la Fuente Sánchez, 2011), en gran medida esto se debe a las sugerencias incluidas en la última edición del manual de APA (2010). Como lo indica ese manual y varios autores (Wilkinson & the Task Force on Statistical Inference, 1999; Wright, 2003; García García et al., 2011, entre otros), estos indicadores ofrecen una interpretación de los resultados más completa que la simple salida dicotómica de una prueba de hipótesis. Además, responden satisfactoriamente a varias de las críticas que se acumulan contra las pruebas de hipótesis (Meehl, 1978; Cohen, 1994; Nickerson, 2000). Las medidas de tamaño del efecto son estandarizadas, por lo que superan el inconveniente de las pruebas de hipótesis en cuanto a su dependencia en el tamaño de las muestras y sirven para comparar investigaciones, una operación muy valiosa en el

terreno de los estudios meta-analíticos, que resumen hallazgos de investigaciones en un determinado dominio.

La literatura sobre este tópico reconoce dos familias de medidas del tamaño del efecto (Kirk, 2003; Ellis, 2010); la de las diferencias entre grupos y las de asociación. El primer objetivo de este documento es reducir estas dos familias a un solo conjunto de indicadores completamente comparables para cuantificar de manera unívoca la magnitud de una asociación entre dos variables o de una diferencia entre grupos, midiendo los tamaños del efecto como la proporción de varianza explicada.

El segundo objetivo del documento es la estimación por intervalo del tamaño del efecto. Es un resultado que se encuentra rara vez en los reportes de investigación, que suelen limitarse a informar el tamaño del efecto como medida descriptiva, obtenida de los datos de la muestra. Sin embargo, para dar valor general a lo que se observa en la muestra, es necesaria una operación de inferencia, como sucede con todos los estimadores descriptivos. En este caso, la inferencia se realiza por medio de intervalos de confianza, incluidos también en las recomendaciones recientes de APA (2010).

Se discute primero la importancia de comunicar de manera comparable los resultados de investigación en Ciencias Sociales, para luego ejemplificar diferentes casos de aplicación de medidas de tamaño del efecto. Esto se hace sobre datos de observación provenientes de haber aplicado una prueba de desarrollo infantil (la escala de Bayley, 1969) a una muestra de 543 niños de entre 0 y 24 meses de edad en la ciudad de Córdoba (Rodríguez, Calderón, Cabrera, Ibarra, Moya y Faas, 2005). A continuación se ilustra la razón por la que la estimación del tamaño del efecto no es una extensión intuitiva de los intervalos de confianza usuales y se muestra el procedimiento para hacerla. Finalmente, se usa el paquete MBESS (Kelley, 2007) en lenguaje R a través de INFOSTAT® (Di Rienzo, Casanoves, Balzarini, Gonzalez, Tablada, Robledo 2012) para realizar las estimaciones de manera sencilla. Parte de lo que aquí se presenta puede hallarse en Busk & Serlin (1992), Steiger & Fouladi (1997), y Cumming & Finch (2002); sin embargo, la aplicación de Infostat® potenciado por su vinculación con R, constituye un aporte novedoso porque otorga accesibilidad a procedimientos que de otra manera son engorrosos o requieren varios pasos.

Comunicación del tamaño del efecto.

Las consecuencias de un tratamiento, de un estilo de vida, de una decisión, de una política, de una catástrofe, de una innovación, etc. sobre algún aspecto de interés (variable de respuesta), pueden expresarse como diferencia entre grupos (comparando los que sufrieron la intervención o que fueron expuestos a diferentes niveles de un tratamiento con quienes no lo hicieron) o como los efectos de una variable (antecedente) sobre otra (consecuente). Estas dos maneras de plantear el problema dan lugar a diferentes métodos para medir la magnitud de los efectos. Si se piensa en términos de comparación de grupos, el problema es cuantificar las diferencias y volver estándar esas cantidades, para poder compararlas con otros efectos; esta es la *familia de los d*, según Ellis (2010). Cuando se plantea el problema como relación entre variables, se generan medidas de la intensidad de la asociación; *la familia de los r* (Ellis, 2010), que pueden transformarse en los coeficientes generales de determinación (por ejemplo R^2 o η^2) que indican qué parte de la variabilidad total de la variable de respuesta es atribuible a las diferentes variables explicativas; se trata de la *proporción de la varianza explicada*. A continuación se verá la forma de transformar las medidas de diferencia entre grupos para que puedan expresarse también como proporción de varianza explicada, de modo que se agregue, a la ventaja de su expresión estandarizada, una lectura llana y comprensible, adecuada para quienes carecen de experiencia en el uso de vocabulario técnico, ya que el uso de esta medida facilita la comunicación de resultados a públicos no especializados.

Las medidas de tamaño del efecto son indicadores descriptivos; dan cuenta de lo observado a nivel de la muestra, pero carecen de validez inmediata para la población general. Una vez que un estudio dio un resultado que se expresa en términos de tamaño de un efecto, es necesario construir el intervalo de confianza que estime su verdadero valor paramétrico. Del mismo modo en que todo hallazgo muestral, requiere de un paso de inferencia para alcanzar validez más allá de los casos observados. Específicamente, para analizar si dos tamaños del efecto difieren significativamente se requiere comparar sus intervalos de confianza, solo en el caso que éstos sean disjuntos, podrán considerarse efectos significativamente diferentes.

El tamaño del efecto como proporción de varianza explicada.

1. Comparación de dos grupos

A fin de explorar el posible efecto de la educación de la madre en el nivel de desarrollo motor de los niños, se comparan los puntajes de la subescala mental del test de Bayley entre niños varones y mujeres. La hipótesis nula de la prueba es que no hay diferencias en el puntaje de la escala mental de los niños de los dos grupos.

H_0: μ_1-μ_2=0 ó bien H_0: $\Delta\mu = 0$

Solicitamos la prueba y obtenemos (tabla 1):

Tabla 1. Salida de Infostat: Prueba T para muestras Independientes Bayment X Sexo.

Clasificación	Variable	Grupo 1	Grupo 2	n (1)	n (2)	Media (1)	Media (2)	t	p-valor
Sexo	Bayment	{2.00}	{1.00}	252	263	99.13	94.99	3.33	0.0009

Nota. Elaborado sobre la base *bayley.idb2* (Rodriguez et al, 2005)

La variable sexo tiene categorías 1=varón 2=mujer, la muestra tiene 252 mujeres y 263 varones, las primeras promedian 99.13 puntos en la escala mental de la prueba de Bayley y los segundos 94.99. En este caso el puntaje t se calculó con

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\Delta\bar{x} - \Delta\mu_0}{s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

En la que $\Delta\mu_0$ es el valor hipotético de $\Delta\mu$ que vale cero bajo la hipótesis nula y s_p es la desviación estándar ponderada[1], que se obtiene a partir de las varianzas de las dos muestras:

$$s_p = \sqrt{\frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}}$$

Debido a que el valor hallado de p es menor a 0.05 (y también menor a 0.01), el resultado que se obtiene de esta prueba es que hay evidencia para rechazar la hipótesis nula, es decir, para rechazar que las medias sean iguales. Si las medias poblacionales fuesen

[1] Este es el caso de considerar iguales a las varianzas poblacionales. El cálculo de la varianza de la diferencia y de los grados de libertad se modifica si este supuesto no se cumple, en cuyo caso se debe aplicar la corrección de Welch Sattewhaite, que INFOSTAT® realiza de manera automática si así lo indica el resultado de la prueba de homogeneidad de varianzas.

iguales, habría una probabilidad de 0.0009 de hallar una diferencia muestral como la observada o más extrema que ella. Esta es la lectura usual del resultado dicotómico de la prueba de hipótesis.

Nuestro interés sigue por el camino de cuantificar esa diferencia, si concluimos que los puntajes difieren según el sexo de los niños, se requiere conocer cuál es el tamaño de ese efecto o la magnitud de esa diferencia, expresada de un modo que sea comparable con otros estudios. Esto puede responderse de manera directa a través de la diferencia estandarizada de las medias de los dos grupos, o bien de manera indirecta tratando el problema como la relación entre dos variables: sexo y puntaje en la subescala. Para la primera, la diferencia se expresa estandarizada a través del coeficiente d de Cohen (1962), el delta de Glass (1976) y el g de Hedges (1981)[2]. En todos los casos se parte de la diferencia de las medias muestrales, pero mientras el de Cohen (y el de Hedges) la dividen en la desviación estándar combinada[3]:

$$d_c = \frac{\bar{x}_1 - \bar{x}_2}{s_p} = \frac{\Delta\bar{x}}{s_p}$$

El de Glass —indicado para diseños experimentales—, usa como denominador a la desviación estándar del grupo control:

$$d_g = \frac{\bar{x}_1 - \bar{x}_2}{s_c} = \frac{\Delta\bar{x}}{s_c}$$

En este ejemplo, corresponde que se use la medida de Cohen, ya que, al tratarse de datos de observación no hay un grupo experimental y uno control. El valor del d de Cohen es 0.293 y se considera (Cohen, 1992) un efecto pequeño. La lectura es que las dos medias difieren en 0.29 veces la desviación estándar promedio[4] de los grupos.

Para expresar este efecto como proporción de la varianza explicada, se debe comparar la varianza que resulta luego de la agrupación en las dos categorías de la variable

---

[2] Aquí es necesario acordar un cambio en la notación habitual. Llamaremos d_c y d_g a las medidas de Cohen y Glass respectivamente, y reservaremos la letra δ (delta) para los parámetros correspondientes (δ_c y δ_g)

[3] El denominador propuesto por Cohen originalmente es $s_p = \sqrt{\frac{\sum(x_{i1}-\bar{x}_1)^2+\sum(x_{i2}-\bar{x}_2)^2}{n_1+n_2-2}}$ y el de Hedges $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$. Si las varianzas de los grupos se calculan con denominador *n-1* (como es lo habitual para evitar el sesgo en la estimación), entonces los denominadores no difieren y la medida es la misma.

[4] En efecto, s_p es un promedio ponderado de las desviaciones estándar de los dos grupos

antecedente, con la varianza total[5]. La primera se estima con la suma de los cuadrados ponderados de los desvíos de las medias de los grupos a la media general, que es 97.02:

$$SCExplicada = (\bar{y}_1 - \bar{\bar{y}})^2 * n_1 + (\bar{y}_2 - \bar{\bar{y}})^2 * n_2$$

La suma de cuadrados de la variable dependiente es

$$SCTotal = {s_y}^2 * n$$

El cociente de estas dos cantidades es 0.021, cuya lectura es que poco más del 2% de las diferencias en los puntajes de la subescala mental se explica por el sexo de los evaluados.

La mayor facilidad para la interpretación de esta medida por sobre el d de Cohen podría debilitarse por la mayor dificultad de su cálculo, sin embargo, se puede llegar a él de una manera más directa. Para ello, se codifican como cero y uno las dos categorías de la variable sexo (las que definen los grupos que se comparan) y se calcula un coeficiente r de Pearson entre la variable así codificada y el puntaje en la prueba. El coeficiente se denomina coeficiente de correlación punto biserial, que en este ejemplo vale 0.14. El punto de mayor interés es que el cuadrado de este coeficiente es el que se conoce como coeficiente general de determinación, que mide la proporción de la varianza total que es explicada por la variable de agrupación, al que se indica R^2. En efecto, el resultado de elevar al cuadrado el coeficiente punto biserial es R^2=0.021, el mismo que se llega por el cociente de las sumas de cuadrados explicada y total.

Alternativamente, puede llegarse a este coeficiente transformando el d de Cohen por medio de:

$$R^2 = \frac{d^2}{d^2 + 4}$$

Además, tiene relación directa con la prueba de diferencia de medias, ya que:

$$R^2 = \frac{t^2}{t^2 + gl}$$

Cualquiera sea el procedimiento elegido para llegar a él, este coeficiente permite una

[5] Para simplificar las expresiones trabajaremos con las sumas de cuadrados, ya que los denominadores son iguales y se

comunicación más clara que la de d, al afirmar que la diferencia de sexos entre los niños da cuenta del 2% de las diferencias en el puntaje de la prueba. Por el contrario, al reportar el índice d de Cohen se indica que la diferencia estandarizada en los puntajes de varones y mujeres es de 0.29, lo que se debe interpretarse como pequeña.

2. Más de dos grupos.

Cuando la variable de agrupación tiene más de dos categorías, se reemplaza la prueba t por un análisis de la varianza de una vía. Es una situación más simple que la anterior para el cómputo de la proporción de la varianza explicada, porque la salida de esta prueba ya contiene las sumas de cuadrados, entonces el cálculo del coeficiente R2 es inmediato; además, es frecuente que los programas de análisis de datos lo incluyan en la salida. Elegimos como ejemplo, sobre la misma base, la medición del efecto de estrato socioeconómico del hogar sobre el puntaje en la subescala mental. La salida del análisis de la varianza es (tabla 2):

Tabla 2. Salida Infostat: Análisis de la varianza para el puntaje de la escala mental como función del estrato socioeconómico.

F.V.	SC	gl	CM	F	p-valor
estrato	3520.23	8	440.03	2.21	0.025
Error	100714.61	506	199.04		
Total	104234.84	514			

Nota. Elaborado sobre la base bayley.idb2 (Rodríguez et al, 2005)

La lectura del valor p conduce al rechazo de H0, por lo que se concluye que los grupos no son iguales. La cuantificación de este efecto se realiza de manera análoga al anterior:

$$R^2 = \frac{SCExplicada}{SCTotal} = \frac{3520.23}{104234.84} = 0.034$$

cancelan al dividirlos.

Que indica que un 3,4% de las diferencias de puntaje en la subescala motora se explican por el estrato socioeconómico del hogar del niño.

3. Más de una variable explicativa.

Cuando se trata con dos o más factores explicativos, se evalúa por separado el efecto de cada uno y el efecto conjunto. Ilustramos esto con el análisis simultáneo de las dos relaciones anteriores (tabla 3):

Tabla 3. Salida de Infostat: Análisis de la varianza para el puntaje de la escala mental como función del estrato socioeconómico y el sexo.

F.V.	SC	gl	CM	F	p-valor
Modelo.	5901.21	9	655.69	3.37	0.0005
Estrato	3520.23	8	440.03	2.26	0.0222
Sexo	2380.98	1	2380.98	12.23	0.0005
Error	98333.63	505	194.72		
Total	104234.84	514			

Nota. Elaborado sobre la base *bayley.idb2* (Rodríguez et al, 2005)

Cuando se trata de más de una variable, el coeficiente R^2 mide el porcentaje de la varianza que es explicado por el conjunto de factores, que es la suma de los efectos que aporta cada uno. El cálculo de las contribuciones de los factores se realiza nuevamente con el cociente de las sumas de cuadrados[6]:

$$R^2_{estrato} = \frac{SCExplicada_{estrato}}{SCTotal} = \frac{3520.23}{104243.84} = 0.034$$

$$R^2_{sexo} = \frac{SCExplicada_{sexo}}{SCTotal} = \frac{2380.98}{104243.84} = 0.023$$

Cuya suma es el coeficiente R^2 y la medida de la parte de la varianza total que es

[6] En la literatura es frecuente denominar a este coeficiente eta cuadrado (η^2) pero no es necesario diversificar la notación para medidas que dan cuenta del mismo concepto: la proporción de la variabilidad total de la que da(n) cuenta una o un

explicada por las dos variables simultáneamente. En el caso que se haya incluido en el ANOVA a la interacción entre los factores explicativos, puede calcularse también la proporción de varianza que explica ese componente. En este ejemplo, el efecto de la interacción no fue significativo.

Estimación por intervalo de la proporción de la varianza explicada.

Hasta este punto se calculó el coeficiente R^2 para diferentes casos como medida descriptiva del tamaño del efecto, el paso siguiente será generalizarlo a la población de la cual provienen las muestras. Se trata de estimar estos tamaños del efecto para la población, por medio de intervalos de confianza, pero estos intervalos no se construyen con el mismo procedimiento que los que estiman la media, la proporción, la diferencia de medias, etc. Veremos por qué la lógica de los intervalos no es igual y luego el modo de construirlos.

La distribución no central de $\Delta\bar{x}$ bajo H_1.

En el primer ejemplo se habría podido construir un intervalo de confianza para estimar la diferencia de medias (no estandarizada) usando la misma distribución t de la prueba de hipótesis, despejando $\Delta\mu$ de la expresión del puntaje estandarizado t y fijando los valores de t, con 513 grados de libertad, de modo de delimitar una confianza del 95%.

$$(\bar{x}_1 - \bar{x}_{2)} \pm t_{n_1+n_2-2,\ 0.025} * s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 4.14 \pm 1.96 * 1.24$$

Los límites resultan 1.70 y 6.58, que indica que el 95% de los intervalos que se construyan extrayendo muestras aleatorias de los dos grupos generará límites que contengan a la verdadera diferencia que existe entre las medias de las dos poblaciones. Se estima así la diferencia bruta, sin estandarizar. Por el contrario, cuando nos interesamos por el tamaño del efecto, el supuesto de $\Delta\mu = 0$ ya no es válido, porque

---

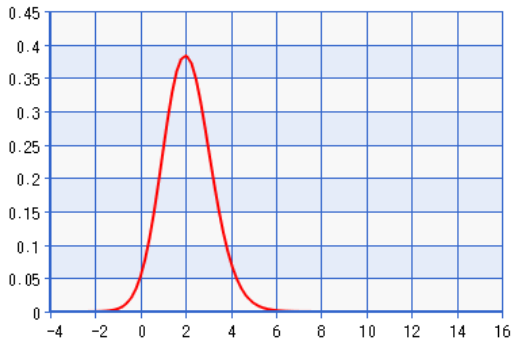
conjunto de variables explicativas.

se ha rechazado la H_0 que lo sostiene, cuya consecuencia es que la distribución t que venimos usando ya no es adecuada. Será entonces necesario usar una distribución que se llama *t no central*, que es semejante a la distribución t de Student, pero que no es simétrica y no está centrada en cero sino en otro punto, al que se denomina *parámetro de no-centralidad*. La explicación de por qué sucede esto es que todo el razonamiento de la prueba de hipótesis se realiza bajo el supuesto según el cual H_0 es verdadera, pero la estimación del tamaño del efecto se realiza cuando se ha rechazado H_0, es decir cuando se ha concluido que existe algún efecto significativo y queremos precisar su magnitud[7]. Por lo tanto, ya no es cierto que la media de la población sea la hipotética ($\Delta\mu = 0$), sino que es otra a la que no conocemos; la llamaremos $\Delta\mu_1$.

Veamos algunos ejemplos de la forma de la distribución t no central[8] que al igual que t, es una familia de distribuciones (figura 1).

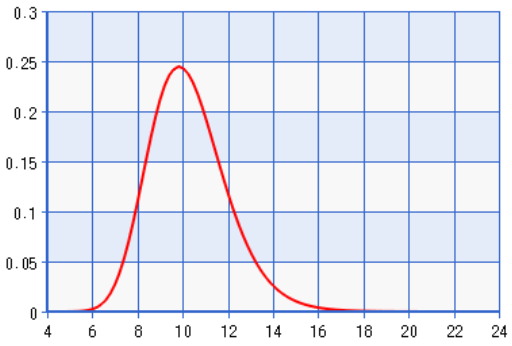
Gl=30, pnc=2

$t_{30,2}$



Gl=30, pnc=10

$t_{30,10}$



[7] Kelley (2007) usa un enfoque diferente para fundamentar el carácter no-pivotal de las medidas de tamaño del efecto y, en consecuencia de la necesidad de construir intervalos asimétricos, pero alcanza la misma conclusión.

[8] Estos gráficos pueden obtenerse del sitio http://keisan.casio.com. La distribución t que conocemos es un caso particular de esta distribución, con parámetro de no centralidad igual a cero.

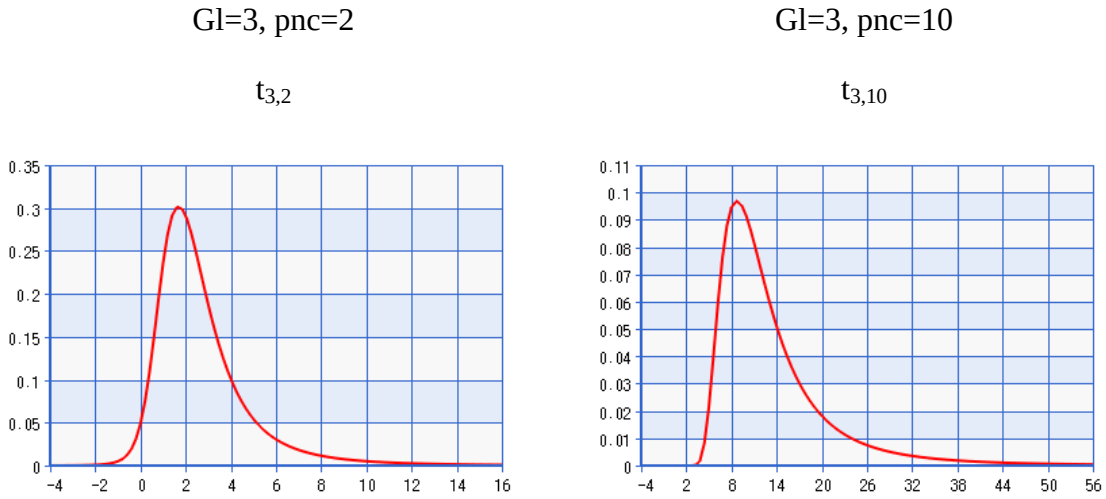


Figura 1. Formas de la distribución t no central con diferentes grados de libertad (Gl) y parámetro de no centralidad (pnc).

Como muestran los gráficos, es una familia de distribuciones sesgadas, cuya forma, tanto en curtosis como asimetría, depende de los grados de libertad y del parámetro de no centralidad; estos dos números determinan en cada caso, cuál es la distribución.

Parámetro de no centralidad y límites del IC para dc.

Veamos por qué aparece el sesgo en la distribución t cuando H_0 se rechaza. La diferencia de medias poblacionales hipotética es cero y llamamos $\Delta\mu_1$ a la verdadera diferencia de medias poblacionales (que desconocemos). Si sumamos y restamos la verdadera media $\Delta\mu_1$ a la expresión original del puntaje estándar en la prueba de diferencia de medias y reordenamos los términos, obtenemos:

$$\frac{\Delta\bar{x} - \Delta\mu_0 + \Delta\mu_1 - \Delta\mu_1}{s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\Delta\bar{x} - \Delta\mu_1}{s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} + \frac{\Delta\mu_1 - \Delta\mu_0}{s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

El primer término del segundo miembro es una distribución t central alrededor de la verdadera diferencia de medias de la población. El término que se le suma

desplaza su centro y ese desplazamiento es el que se llama *parámetro de no-centralidad*. Está formado por la distancia entre la diferencia de medias verdadera ($\Delta\mu_1$, desconocida), y la hipotética ($\Delta\mu_0$ que vale cero), expresada en términos del error estándar de la diferencia:

$$pnc = \frac{\Delta\mu_1}{s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Que resulta ser casi lo que buscamos como tamaño del efecto poblacional, porque éste es esa misma diferencia solo que expresada en términos de la desviación estándar muestral:

$$\delta_c = \frac{\Delta\mu_1 - \Delta\mu_0}{s_p}$$

O simplemente

$$\delta_c = \frac{\Delta\mu_1}{s_p}$$

Entonces la relación entre el parámetro de no centralidad y el tamaño del efecto es:

$$\delta_c = pnc * \sqrt{\frac{n_1 + n_2}{n_1 * n_2}}$$

Ahora el problema de estimar δ_c se reduce a calcular los límites para el parámetro de no centralidad. El valor t hallado en la prueba, de 3.33 servirá para responder a las siguientes preguntas:

¿Cuál es la distribución t (no central) que deja una probabilidad extrema inferior de 2.5% para el valor 3.33?

¿Cuál es la distribución t (no central) que deja una probabilidad extrema superior de 2.5% para el valor 3.33?

Identificar a una distribución t no central implica conocer sus grados de libertad y su parámetro de no-centralidad. Los primeros dependen del tamaño de la muestra, en el caso de la prueba de diferencia de medias son simplemente $n_1 + n_2 - 2$. El problema es entonces hallar los parámetros de no-centralidad de esas distribuciones. Reformuladas, las preguntas se reducen a ¿Cuáles son los parámetros de no-centralidad de las distribuciones de la media muestral para las que el puntaje t observado es el percentil 2,5 y 97,5? Formalmente,

preguntamos por los dos *pnc* que cumplen que:

$$P\left(t_{513,pnc} < 3.33\right) = .025$$

$$P\left(t_{513,pnc} < 3.33\right) = .975$$

Hay más de un modo de encontrar estos valores, una opción (Howell, 2011) es de manera reiterada, procediendo por aproximaciones sucesivas. En este caso lo usual es trabajar con un software como R, iniciar la búsqueda con un valor e ir refinando sucesivamente, hasta alcanzar un valor aceptablemente próximo a las probabilidades fijadas. Este proceso conduce a encontrar[9]

pt(3.33, 513, 1.4)= 0.9724

pt(3.33, 513, 5.3)= 0.0249

Que son aproximaciones que podrían mejorarse continuando con la iteración. Sin embargo, resulta más directo usar una calculadora de probabilidades que puede hallarse on line, como la de keisan casio (http://keisan.casio.com/exec/system/1234508566), a la que se ingresa con la probabilidad acumulada (0.025), el valor t (3.33) y los grados de libertad (513), para encontrar el parámetro de no centralidad de la distribución correspondiente. Repitiendo la operación con la probabilidad acumulada de 0.975, se encuentra el pnc de la otra distribución. Con este procedimiento se encuentran:

$$P\left(t_{513,1.3579} < 3.33\right) = .975$$

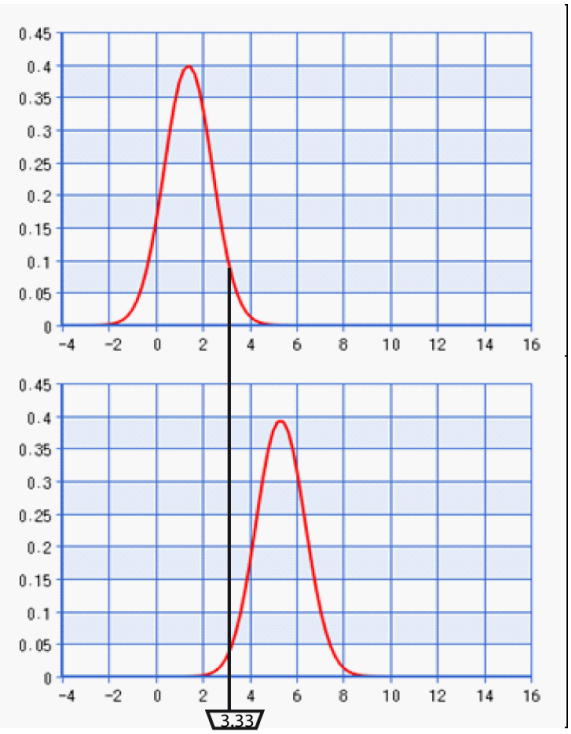
$$P\left(t_{513,5.2989} < 3.33\right) = .025$$

Que son cercanos a los hallados por aproximación usando R. Será suficiente tomar 1.36 y 5.30 para los dos parámetros de no centralidad

Con estos dos resultados encontrados, disponemos de un intervalo de confianza del 95% para el parámetro de no-centralidad de la distribución de las diferencias de medias muestrales. El intervalo es [1.36; 5.30]. Las dos distribuciones halladas son (figura 2):

[9]La que sigue es la sintaxis de R para solicitar probabilidades bajo la distribución t, el primer número es el puntaje t, el segundo los grados de libertad y el tercero es el que se alcanza por reiteración hasta aproximar las probabilidades buscadas (0.975 y 0.025)

$t_{513,\ 1.36}$



$t_{513,\ 5.30}$

Figura 2. Distribuciones t no centrales con 513 grados de libertad y parámetros de no centralidad correspondientes a los percentiles 95 y 5 del puntaje t (central)=3.33.

De este modo, los valores 1.36 y 5.30 son los *pnc* de las dos distribuciones que buscamos.

Con ellos podemos calcular los límites del intervalo de confianza que estima δ_c:

$$Li = 1.36 * \sqrt{\frac{263 + 252}{263 * 252}} = 0.12$$

$$Ls = 5.30 * \sqrt{\frac{263 + 252}{263 * 252}} = 0.47$$

Tenemos entonces los límites del intervalo de confianza que estima el verdadero efecto a nivel de la población, medido a través del coeficiente d de Cohen. La relativa complejidad de este procedimiento es lo que limita la difusión de su uso, sin embargo, la estimación por intervalo de las medidas de tamaño del efecto es muy valiosa porque autoriza las comparaciones a escala poblacional. En efecto, la diferencia entre dos tamaños del efecto será significativa si los intervalos que las estiman son disjuntos; por el contrario, si hay superposición entre ellos, la diferencia descriptiva que se encuentre no será reflejo de una

verdadera diferencia a nivel de la población y esto es muy importante de comunicar cuando se reportan resultados de investigación.

INFOSTAT y R para obtener el intervalo.

Una vez expuesto el procedimiento que conduce a determinar los límites del intervalo de confianza para estos tamaños del efecto, podemos hacer uso de un paquete específico que está disponible en lenguaje R. El paquete se denomina MBESS: Métodos para la Ciencias del Comportamiento, Educativas y Sociales (Kelley et al, 2011) e incorpora un conjunto de rutinas, entre ellas las que sirven para la estimación de tamaños de efecto basados en distribuciones no centrales. Una limitación de R es que resulta engorroso para usuarios no especializados, en especial en la manipulación de los datos, lo que puede subsanarse realizando esas operaciones con algún paquete conocido y dejando para R solo los procedimientos de mayor complejidad. Esto se ve muy facilitado por la vinculación entre INFOSTAT ® y R, medio por el que se trabaja de manera usual en INFOSTAT® para luego cargar la base activa y operar sobre ella con los comandos de R. Veamos esto aplicado al ejemplo actual.

Comenzamos corriendo la prueba t en INFOSTAT para el primer ejemplo, que está mostrada en la Salida 1. Para generar el intervalo de confianza para d de Cohen será necesario conectar INFOSTAT con R[10] y acceder a él por el ícono [R] en la barra de menú. Allí se carga el paquete MBESS escribiendo la instrucción:

install.packages("MBESS")

Que pedirá que se elija desde qué sitio realizar la descarga, se recomienda optar por "cloud" (la nube) ya que así se selecciona el sitio espejo de manera automática. Luego se instala el paquete desde la biblioteca indicando:

library("MBESS")

[10]En el comando *Ayuda* aparece la opción *¿Cómo instalar R?* Una vez cumplidos los pasos que allí se indican, se debe solicitar *Intentar comunicación con R*, también dentro del menú *Ayuda* y aparecerá el ícono de [R] en la barra de menú. Solo debe hacerse esto una vez.

Este paquete contiene las rutinas necesarias para hacer los intervalos de confianza de tamaños del efecto en muy variadas situaciones. El correspondiente al d de Cohen se solicita con el comando ci.smd (confidence interval standard mean difference) al cual se debe informar: el parámetro de no centralidad inicial, que es el valor absoluto del puntaje t, los tamaños de muestra y el nivel de confianza. En el ejemplo de la tabla 1 es:

ci.smd(ncp=3.33, n.1=263, n.2=252, conf.level=.95)

Y se obtiene:

$Lower.Conf.Limit.smd *[Límite inferior del intervalo de confianza]*

[1] 0.1196973

$smd

[1] 0.2935416 *[Diferencia de medias estandarizada, d de Cohen]*

$Upper.Conf.Limit.smd

[1] 0.4671028 *[Límite superior del intervalo de confianza]*

Que son los mismos resultados que se habían obtenido de manera manual. La interpretación sigue las mismas reglas de los intervalos de confianza clásicos: El 95% de los intervalos construidos siguiendo este procedimiento contendrá al verdadero valor del tamaño del efecto poblacional, medido por d de Cohen.

A continuación se compara este resultado con la diferencia de puntaje entre hijos de madres con nivel educativo máximo (código 10 en la variable) y mínimo (codificado 1). Debido a que se seleccionan solo hijos de madres con estos niveles educativos, la cantidad de casos en cada grupo es muy diferente a la prueba anterior. La prueba correspondiente arroja (tabla 4):

Tabla 4. Salida de Infostat: Prueba t para muestras Independientes Grupos: niveles educativos extremos de la madre.

Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)	t	p-valor	prueba
Bayment	{0}	{1}	29	12	104.21	89.58	3.09	0.0037	Bilateral

Que indica que la diferencia es significativa. Para cuantificar la diferencia de manera

estándar y poder comprarla con otros estudios, solicitamos al paquete MBESS:

ci.smd(ncp=3.09, n.1=29, n.2=12, conf.level=.95)

Que devuelve el valor de d de Cohen y los límites del intervalo ed confianza del 95%:

$Lower.Conf.Limit.smd

[1] 0.3420741

$smd

[1] 1.060622

$Upper.Conf.Limit.smd

[1] 1.767033

Que nos indica el estimador puntual del efecto; 1.061 y los límites del intervalo de confianza del 95%: [0.34; 1.77]. Observemos que este intervalo se superpone al que estima la diferencia de puntaje mental entre varones y mujeres, de modo que, aunque el estimador puntual del efecto es bastante mayor (1.06 frente a 0.29) no hay suficiente evidencia para sostener que la diferencia entre sexos de los puntajes en la escala mental sea diferente que la que se da entre hijos de madres con nivel educativo dispar.

En el caso de trabajar con un diseño experimental y medir el efecto a través del d de Glass, el comando R para pedir el intervalo a una confianza del 95% es:

ci.smd.c(ncp =XXX, n.C = XXX, n.E = XXX, conf.level = .95)

En la que ncp es, como antes el puntaje t, n.C el número de caso en el grupo control y n.E el número de caso en el grupo experimental. La salida tiene el mismo formato que ci.smd.

MBESS para la proporción de la varianza explicada.

Para la comparación de grupos, hemos visto que la proporción de varianza explicada (R^2) puede calcularse como cociente de sumas de cuadrados explicada y total, o también como el cuadrado del coeficiente punto biserial, ahora nos interesa

construir un intervalo de confianza que lo estime para la población. La rutina del paquete MBESS es general, construida con el vocabulario del análisis de la varianza, por lo que será necesario hacer algunas adaptaciones. El commando es *ci.pvaf* (confidence interval for the proportion of variance accounted for) y la sintaxis:

ci.pvaf(F.value=XX, df.1=XX, df.2=XX, N=XX, conf.level=XX)

Que tiene como insumos:

El puntaje F, F.value, los grados de libertad del numerador y del denominador (df.1 y df.2), el tamaño de la muestra (N) y el nivel de confianza.

Con los datos del ejemplo 1 resulta:

ci.pvaf(F.value=11.08, df.1=1, df.2=513, N=515, conf.level=.95)

El puntaje F se obtiene de la prueba t, elevando al cuadrado el puntaje t, los grados de libertad se interpretan como en el análisis de la varianza, en el numerador siempre es un grado de libertad, porque se comparan dos grupos, en el denominador, *n-1*. El resultado de la operación es:

[1] "The 0.95 confidence limits (and the actual confidence interval coverage) for the proportion of variance of the dependent variable accounted for by knowing group status are given as:"

$Lower.Limit.Proportion.of.Variance.Accounted.for

[1] 0.003560364

$Probability.Less.Lower.Limit

[1] 0.025

$Upper.Limit.Proportion.of.Variance.Accounted.for

[1] 0.05167761

De la que nos interesan los límites del intervalo: 0.0035 y 0.0517, a los que leemos que hay una confianza del 95% que ese intervalo contenga al verdadero valor de la proporción de varianza de los puntajes de la prueba mental, explicada por la diferencia de género.

Cuando se trata de más de dos grupos definidos por una única variable explicativa, el

procedimiento es el mismo: luego de haber realizado el análisis de la varianza (en INFOSTAT®) se solicita, en R, el cálculo del intervalo de confianza. Para el ejemplo tabla 2:

ci.pvaf(F.value=2.21, df.1=8, df.2=506, N=515, conf.level=.95)

Que devuelve los límites 0.000 y 0.054. Si reducimos la confianza al 90%, para mejorar la precisión de la estimación, pedimos:

ci.pvaf(F.value=2.21, df.1=8, df.2=506, N=515, conf.level=.90)

Y los límites son 0.002 y 0.048

Finalmente, una prueba que analiza el efecto de dos variables explicativas, da lugar a una proporción de la varianza explicada por cada una de ellas y a la suma de ellas, que es lo que el modelo completo aporta a la explicación de la variable dependiente. Usaremos la misma rutina de MBESS, por separado para el modelo y para cada variable:

ci.pvaf(F.value=3.37, df.1=9, df.2=505, N=515, conf.level=.95) #Modelo

ci.pvaf(F.value=2.26, df.1=8, df.2=505, N=515, conf.level=.95) #Estrato

ci.pvaf(F.value=12.23, df.1=1, df.2=505, N=515, conf.level=.95) #Sexo

Los puntajes F y los grados de libertad provienen de la Salida 3. El resultado son los intervalos de confianza del 95% para cada caso:

[0.012; 0.083] #Modelo

[0.005; 0.055] #Sexo

[0.000; 0.056] #Estrato

El resultado del análisis puede presentarse indicando el estimador puntual de la proporción de varianza explicada por cada factor y los límites de cada intervalo, para ofrecer una idea cabal de la relación entre las variables. El resumen de esa información tiene la forma:

Tabla 5. Estimación puntual y por intervalo de la proporción de la varianza de los puntajes en la subescala mental explicada por el sexo de los evaluados y el estrato socioeconómico de sus hogares.

	Porcentaje explicado de la varianza total	IC (95%)	
		Li	Ls
Sexo	2.3%	0.5%	5.5%
Estrato	3.4%	0.0%	5.6%
Modelo	5.7%	1.2%	8.3%

Discusión

El reporte de medidas de tamaño del efecto va siendo parte de la rutina de investigación en Ciencias Sociales, sin embargo, la posibilidad de hacer comparaciones, que es su virtud principal, se ve limitada por dos razones. Una razón es que hay una diversidad de medidas de tamaño del efecto cuya lectura no es idéntica y que, aunque en general sean reductibles unas a otras, no siempre los reportes de investigación ofrecen todos los datos necesarios para calcular medidas diferentes de las comunicadas. La otra razón es que la práctica es informar solo medidas descriptivas del tamaño de efecto y no son frecuentes las estimaciones a la población que permitan decidir si dos resultados difieren o no, comparando sus intervalos de confianza.

En este trabajo se propone resolver la primera limitación por medio del uso de una medida general de tamaño del efecto: la proporción de la varianza explicada por la(s) variable(s) antecedente(s). La ventaja de esta medida es que puede calcularse sin dificultad para los análisis que implican comparaciones entre grupos, y que son inmediatas en los planteos de relaciones entre variables. Además, no es necesario que el lector conozca valores de referencia para interpretar su significado.

La segunda limitación mencionada se supera construyendo intervalos de confianza para estimar tamaños del efecto estandarizados. Estos no suelen ser tratados en detalle en los manuales sobre estadística en Ciencias Sociales debido a que su cómputo es complejo, a menudo solo disponible a través de scripts especializados en paquetes de análisis de datos, y

por eso restringido a investigadores que hacen uso intensivo de técnicas estadísticas. Sin embargo, aquí se ha recurrido al paquete MBESS que vuelve accesible la construcción de estos intervalos. Aunque MBESS sea un paquete en lenguaje R, esto no implica habilidades de programación por parte del usuario ni experiencia en R, ya que se accede a él por medio de INFOSTAT. Así, el análisis de datos se realiza de manera usual en INFOSTAT® (como se lo podría hacer en cualquier paquete de análisis de datos) y luego se pasa a R solo para la construcción de los intervalos de confianza con el paquete MBESS.

Se considera que expresar los efectos que produce una intervención como un intervalo de confianza que estima la proporción de la varianza explicada en la población, es una manera clara y comparable de informar resultados de investigación, por lo que constituye un aporte a la acumulación de conocimiento proveniente de investigación en Ciencias Sociales.

Recomendamos la consulta al documento *Package 'MBESS'*, disponible en http://cran.r-project.org/web/packages/MBESS/MBESS.pdf, en el que se describen todas las rutinas del paquete, su sintaxis y los insumos que cada una requiere para aplicarse.

Referencias

AERA [American Educational Research Association] (2006). Standards on reporting on empirical social science research in AERA publications. *Educational Researcher,* 35, 33–40.

Algina, J., Keselman, H. J., & Penfield, R. D. (2006).Confidence interval coverage for Cohen'seffect size statistic. *Educational and Psychological Measurement*, 66, 945-960.

APA [American Psychological Association]. (2010). *Publication manual of the American Psychological Association* (6ta ed.). Washington, DC

Cohen, J. (1962). "The statistical power of abnormal-social psychological research: A review" *Journal of Abnormal and Social Psychology*, 65(3): 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist* 49(12), 997–1003.

Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 633-649.

Di Rienzo J.A., Casanoves F., Balzarini M.G., Gonzalez L., Tablada M., Robledo C.W. (2012) *InfoStat versión 2012*. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar

Field, A. (2013). Effect sizes. Documento metodológico en *Statistics Hell*. Disponible en http://www.statisticshell.com/docs/effectsizes.pdf accedida 3/7/2013

Frías Navarro, M., Llobell, J. & García Pérez, J. (2000). Tamaño del efecto del tratamiento y significación estadística *Psicothema* Vol. 12, Suplem.2, pp. 236-240

García García, J.; Ortega Campos, E.; De la Fuente Sánchez, L. (2011). The Use of the Effect Size in JCR Spanish Journals of Psychology: From Theory to Fact. *The Spanish Journal of Psychology*, 1050-1055.

Glass, G. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher* 10: 3-8.

Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators *Journal of Educational Statistics*, 6(2): 106–128.

Howell, D. (2011). Confidence Intervals on Effect Size. Suplemento web de *Statistical Methods for Psychology*. Cengage Learning. Disponible en http://www.uvm.edu/~dhowell/methods8/Supplements/Confidence%20Intervals%20on%20Effect%20Size.pdfaccedida 3/7/2013

Kelley, K. & Lai, K. (2011). MBESS: MBESS. R package version 3.2.1.http://CRAN.R-project.org/package=MBESSaccedida 3/7/2013

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, 65, 51-69.

Kelley, K. (2007). Methods for the Behavioral, Educational, and Social Sciences: An R package *Behavior Research Methods*, 39 (4), 979-984

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, 11(4), 363–385.

Kirk, R.E. (2003). The importance of effect magnitude. En S.F. Davis (editor), *Handbook of Research Methods in Experimental Psychology*. Oxford, UK: Blackwell.

Lai, K. & Kelley, K. (2013).The MBESS Package. *R wiki*. Disponible en http://rwiki.sciviews.org/doku.php?id=packages:cran:mbessaccedida 3/7/2013

Meehl, P. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology, *Journal of Consulting and Clinical Psychology* Vol, 46, 806-834

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy, *Psychological Methods*, 5, 241-301.

Rodríguez, M., Calderón, L., Cabrera, L., Ibarra, N., Moya, P. y Faas, A. E. (2005). Análisis de Consistencia Interna de la Escala Bayley del Desarrollo Infantil para la Ciudad de Córdoba (Primer año de Vida). *Revista Evaluar*. Laboratorio de Evaluación Psicológica y Educativa. Facultad de Psicología Nº 5 Universidad Nacional de Córdoba

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, 61, 605-632.

Smithson, M.J. (2003). Confidence Intervals. *Quantitative Applications in the Social Sciences Series*, No. 140. Thousand Oaks, CA: Sage.

Steiger, J. & Fouladi, R. (1997). Noncentral interval estimation and the evaluation of statistical models. En L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds).*What if there were no significance tests?* Mahwah, N. J., Lawrence Erlbaum Associates.

Steiger, J. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods,* 9, 164-182.

Sun, S., Pan, W., & Wang, L. (2010). A Comprehensive Review of Effect SizeReporting and Interpreting Practices in Academic Journals in Education and Psychology.*Journal of Educational Psychology*. Publicación on line.doi: 10.1037/a0019507

Wilkinson, L. & the Task Force on Statistical Inference (1999). Statistical Methods in Psychology Journals Guidelines and Explanations. *American Psychologist* American Psychological Association. Inc. Vol. 54, No. 8, 594-604

Wright, D. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, 73, 123–136. The British

Psychological Society. University of Sussex, UK

Bayley, N. (1969). *Manual for the Bayley scales of infant development*. California, EE.UU.: The Psychological Corporation.

Cohen, J (1992). A power primer. *Psychological Bulletin* 112 (1): 155–159.