EVALUAR

Laboratorio de Evaluación Psicológica y Educativa

Evaluar, 11 (2012), 59 – 84
ISSN 1667-4545

Tendencias en el análisis estadístico: Límites de la inferencia frecuencial y posibilidades del enfoque bayesiano

Eduardo Bologna[1]

Facultad de Psicología, Universidad Nacional de Córdoba - Universidad Católica de Córdoba. Argentina.

Resumen. Los procedimientos estadísticos que se usan en investigación en Psicología se hallan en plena revisión. La acumulación de críticas hacia las técnicas clásicas ha suscitado respuestas diferentes. Por un lado, profundizar los análisis más allá de las pruebas de hipótesis y dar más detalles a la lectura de los datos. Por otro lado, cada vez son más los partidarios de introducir el enfoque bayesiano al hacer inferencias y crecen las rutinas informáticas para facilitarlo. Sobre el enfoque frecuencial, en este trabajo se analizan los procedimientos adicionales a las pruebas de hipótesis, que sugiere APA en la sexta edición del manual de publicación. Se revisan los fundamentos lógicos de las pruebas de hipótesis, se muestra la utilidad de informar medidas sobre el tamaño del efecto ―junto a la dificultad para inferir sobre ellas―, analizar la potencia, y las diferentes alternativas para calcular intervalos de confianza. Sobre el enfoque bayesiano, se discuten los supuestos y las ventajas en cuanto a la información que ofrece y su aporte al carácter acumulativo de los resultados. Luego se muestra una aplicación simple pero ilustrativa del modo en que razona este enfoque, comparando la estimación de la proporción a través del intervalo de credibilidad y el de confianza.
Palabras clave: Inferencia Frecuencial; Inferencia Bayesiana; Prueba de Hipótesis; Estadística en Psicología

Title: Trends in statistical analysis: The frequentist inference limitations and the possibilities of the Bayesian approach.
Abstract: Statistical procedures used in psychological research are in full review. The accumulation of criticism upon classical techniques has led to different answers. On one hand, there is a deepening on pushing analysis beyond hypothesis testing and more elaborate interpretation of data. On the other hand, more and more researchers suggest introducing Bayesian approach to make inferences, and there are more software routines to facilitate this. On the frequentist approach, this paper discusses additional procedures to test hypotheses suggested in the sixth edition APA Publication Manual. We review the rationale of hypothesis testing; we show the usefulness of reporting effect size measures ―and the difficulty to infer about them―, power analysis, and the different alternatives to calculate confidence intervals. On the Bayesian approach, we discuss the assumptions and the advantages in terms of information it offers and its contribution to the cumulative nature of the results. Then we show a simple illustrative application of its rationale, comparing the estimation of proportions by means of credibility and confidence intervals.
Key Words: Frequency inference; Bayesian Inference; Hypothesis Testing; Statistical Psychology.

1. Introducción.

Paul Meehl, en 1978 decía que las teorías en las áreas "blandas" de la Psicología carecen del carácter acumulativo que es propio del conocimiento científico y una razón que postulaba para esto es la excesiva confianza en las pruebas de hipótesis (NHST)[2]. Más tarde, Schmidt (1996) afirma que los resultados de meta-análisis muestran que la confianza en las pruebas de significación retrasa los procesos de acumulación de conocimiento; y los ataques hacia estos procedimientos se han multiplicado, al punto que en 1999 la APA encargó a un

[1] Por favor dirigir la correspondencia relacionada con este artículo a:
Dr. Eduardo Bologna, ebologna@gmail.com

[2] Cohen (1994) atribuye los primeros ataques a Joseph Berkson, en 1938

grupo de trabajo[3] el análisis de esas críticas y de las propuestas de eliminar las pruebas de hipótesis de las publicaciones en Psicología. Aunque ese grupo no recomendó que se eliminaran las NHST, sí puso énfasis en ampliar el repertorio de procedimientos estadísticos que se usan y fue partidario de una política de inclusión que permita que cualquier procedimiento que arroje luz sobre los fenómenos de interés, se incluya en el arsenal de la investigación científica. La sexta edición del manual (APA, 2010) apoya sólidamente estas recomendaciones y menciona como elementos de ese arsenal: medidas de tamaño del efecto, cálculo de la potencia, intervalos de confianza, profundización del meta análisis. En efecto, la Estadística es una de las siete áreas clave que el grupo de trabajo encargado por la APA, identificó como necesarias de revisión en las recomendaciones para investigadores en Psicología. Como efecto de ello, la sexta edición del manual (APA, 2010) incorpora sugerencias precisas sobre el tema; en particular, el capítulo 4 ofrece una guía para comunicar resultados de inferencias estadísticas. Como señala la cita siguiente, el investigador debe tener un manejo de los procedimientos estadísticos que le permitan ser flexible con las decisiones que toma sobre las herramientas más convenientes para el análisis de sus datos.

> Históricamente, los investigadores han confiado ampliamente en las pruebas de hipótesis como un punto de partida de muchos (aunque no todos) los enfoques analíticos. APA enfatiza que las pruebas de hipótesis no constituyen sino un punto de partida y que, para ofrecer una más completa interpretación de los resultados, es necesario reportar otros elementos, tales como tamaños del efecto, intervalos de confianza y descripciones extensivas. […] Cuando se reporten resultados de inferencias estadísticas o cuando se ofrezcan estimaciones de parámetros o tamaños del efecto, se debe incluir suficiente información para ayudar al lector a comprender completamente los análisis realizados y posibles explicaciones alternativas para los resultados de esos análisis. Dado que cada técnica analítica depende de diferentes aspectos de los datos y de los supuestos, es imposible especificar qué constituye un "conjunto suficiente de estadísticos" para todos los análisis. (APA, 2010, p.33, traducción propia)

Resulta claro que no es posible hacer usos rutinarios de los procedimientos conocidos ni tampoco adoptar técnicas novedosas y fácilmente aplicables desde los paquetes estadísticos. Por el contrario, para analizar con fundamento y creatividad los datos disponibles se requiere un conocimiento que podría denominarse de "usuario competente" que, sin ser el de los expertos en estadística, incluya un repertorio amplio de técnicas, sus fundamentos

[3] Llamado *Task Force on Statistical Inference*. El informe preliminar puede hallarse en: http://www.apa.org/science/tfsi.html

lógicos, sus condiciones de utilización; de tal manera que se logren lecturas precisas de los resultados de su aplicación, que no dejen de lado las limitaciones de las conclusiones a que se llega.

El caso particular de las pruebas de hipótesis y los procedimientos derivados de su forma de razonamiento, son problemáticos porque el resultado es insatisfactorio y porque eso lleva a que sea alto el riesgo de interpretaciones erróneas. El resultado es insatisfactorio porque la conclusión que se alcanza luego de aplicar el test no informa sobre qué tan probable es una hipótesis a la luz de determinados resultados muestrales; por el contrario, indica la probabilidad de ciertos resultados muestrales, si la hipótesis (que llamamos nula) fuera verdadera.

Debido a que lo que interesa al investigador es hacer un juicio sobre sus hipótesis, es muy frecuente caer en errores como considerar que la hipótesis nula se acepta cuando no hay evidencia para rechazarla o interpretar una hipótesis nula rechazada como un hallazgo científico o bien no tener en cuenta el riesgo de error de tipo II. Shrout, en 1997 señalaba que, aunque estos errores habían sido discutidos durante décadas, no hallaba evidencia que hubiese habido impacto en la discusión académica; la última edición del manual de APA (2010) indica que esa dirección empieza a tomarse.

Sin embargo, como lo sugiere Orlitzky (2012), creemos los cambios que serían necesarios para que otras formas de razonamiento estadístico ―como podría serlo la inferencia bayesiana―, ocupen el lugar de las NHST son profundas y no alcanza con la iniciativa de investigadores individuales, sino de reformas promovidas desde las instituciones que rigen los modos admitidos de argumentar estadísticamente.

2. Amplia difusión en el uso del valor p

Muchas publicaciones recientes mantienen un gran respeto hacia las pruebas de hipótesis y al resumen de su resultado a través de valores p. Por ejemplo:

> Este análisis reveló un efecto principal significativo de la dimensión evaluación de modo que los participantes evaluaron su habilidad social ($M = 7.23$, $SD = 1.53$) más favorablemente que su rendimiento en la prueba ($M = 6.22$, $SD = 1.76$), $F(1, 128) = 60.32$, $p < .001$, $r = .57$ [2]. Más aun, hubo un efecto principal significativo de la dirección de la comparación social tal que los participantes ubicados en la condición comparativa inferior ($M = 7.24$, $SD = 1.14$) muestran autoevaluaciones más altas que los participantes en la condición comparativa superior ($M = 6.11$, $SD = 1.50$), $F(1, 128) = 24.05$, $p < .001$, $r = .40$, lo que sugiere

> que la manipulación de la valencia de retroalimentación fue exitosa. (Buckingham et al, 2011, p. 4, traducción propia)
>
> La correlación entre el factor 1 y el factor 4 resultó estadísticamente significativa (r_{14} = -,50, p<,01), al igual que la correlación entre el factor 2 y 4 (r_{24} = -,25, p<,01). El resto de las correlaciones entre pares de factores no resultaron significativos. (Coello y Fernández, 2011, p. 182)

Aún con las crecientes dudas sobre las pruebas de hipótesis (Meehl, 1967, 1978; Cohen, 1994; Haller & Krauss, 2002), estas citas arbitrariamente elegidas entre publicaciones recientes, muestran que el procedimiento conserva amplia difusión en Psicología y que las recomendaciones del manual 2010 de APA son solo parcialmente tenidas en cuenta. Entre estas recomendaciones se cuentan: reportar medidas de tamaño del efecto, indicar el valor exacto de p en lugar de ofrecer una cota máxima, evitar calificar como "significativo" un resultado, porque es confuso en qué sentido lo es, reportar intervalos de confianza, evaluar la potencia. La práctica de identificar valores p menores al 10, al 5 ó al 1%, sigue siendo una rutina con la que se separa lo significativo de lo no significativo que no ayuda a distinguir entre significación estadística y sustantiva. La falta de especificación de la significación entre estadística y clínica o práctica, deja confusos los resultados. Se trata de una distinción de la mayor importancia, porque la significación estadística se refiere a discernir si un resultado observado puede atribuirse al azar, mientras que la significación práctica trata sobre la utilidad del resultado en la realidad (Kirk, 1996).

La fe en las conclusiones a que lleva el rechazo de la hipótesis nula, así como la identificación de *estadísticamente significativo* con *hallazgo científico*, queda bien ilustrada en la existencia de una publicación[4] que recoge resultados de investigaciones que "fracasaron en rechazar H_0", con el objetivo de reducir la creencia sobre resultados no significativos como equivalentes a malos resultados. Los editores de esta revista electrónica señalan que esta creencia conduce a que los investigadores no informen resultados cuando no alcanzan los niveles de significación tradicionales y que, como resultado, las publicaciones contengan solo artículos que alcanzan el nivel de nivel de significación estipulado. Las investigaciones en que no se encontraron resultados significativos se hallan ausentes de la literatura, por eso, sin el recurso de esta publicación, afirman los editores, los investigadores podrían perder su tiempo examinando interrogantes empíricos que ya han sido tratados.

[4] *Journal of Articles in Support of the Null Hypothesis* http://www.jasnh.com/

La lectura del resultado de una prueba de hipótesis es resumida en las publicaciones de manera muy compacta, a través del reporte del valor p. Con una rutina de lectura según la cual si p es menor a cierto valor, se considera que el resultado es "significativo". La facilidad de los programas de análisis de datos, permite hacer una revisión muy rápida de los resultados significativos, ya que adosan, junto al número hallado (diferencia de medias, coeficientes de correlación, etc.) uno, dos o tres asteriscos, indicando que ese número es menor a 0,10, 0,05 ó 0,01. Cuanto más pequeño sea el valor p, parecería inferirse que hay más evidencia para creer que se ha dado con un resultado de interés para la investigación. El riesgo de interpretar resultados significativos como una reducción en la credibilidad de H_0 es muy acentuado. No es difícil deslizar desde "la diferencia entre el grupo experimental y control es significativa, con $p<0,000001$" a afirmar que "hay mucha evidencia para creer que los grupos difieren". El error es grave y lo alertó, no por primera vez, pero sí con gran énfasis y estilo, Jacob Cohen (1994). Para el autor, es tan fuerte la necesidad que tiene el investigador de saber algo acerca de la hipótesis que llega al punto de creer que el valor p puede informarle sobre la plausibilidad de H_0. En ese acto confunde el resultado de la prueba, que es $P(D/H_0)$, con aquello que quisiera saber: $P(H_0/D)$, la probabilidad de la hipótesis dados los datos observados.

El valor p no es más que la probabilidad de hallar un resultado como el observado o más extremo, si H_0 fuera verdadera. Nada dice el valor p acerca de H_0. Fisher (1959) señala que una prueba de significación no nos autoriza a hacer ninguna afirmación sobre las hipótesis en cuestión, en términos de probabilidad matemática. El razonamiento de la prueba de hipótesis, contra lo que puede parecer en una primera aproximación, es deductivo, dado que su punto de partida es suponer un valor poblacional y una distribución de su estimador en todas las muestras posibles extraídas de esa población, para luego deducir la probabilidad de algunos de estos valores (Díaz Batanero, 2007).

3. Efecto del tamaño y tamaño del efecto

Una de las limitaciones que suelen atribuirse a la prueba de hipótesis es su dependencia del tamaño muestral. Una diferencia de medias de un valor absoluto dado puede ser significativa si proviene de una muestra de 300 casos pero no serlo si la muestra solo tiene 30. Este problema dificulta las comparaciones, en particular en el meta-análisis, cuando se comparan resultados provenientes de experimentos que usan diferentes cantidades de casos.

Sin embargo es una consecuencia válida y esperable debido a que hay más certeza en una diferencia que proviene de comprar muchos casos que de pocos.

La salida siguiente proviene de la comparación de dos grupos de 10 observaciones ficticias cada uno. La H_0 afirma que los dos grupos no difieren. Se encuentra una diferencia muestral de 3,8 a la que corresponde un valor p de 0,121, que conduce, según los cánones vigentes, a no rechazar H_0 y afirmar que no hay evidencia para creer que los grupos difieran.

Tabla 1

Prueba t de diferencia de medias, datos ficticios, $n_1=n_2=10$

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)	Var(1)	Var(2)	pHomVar	T	p-valor
grupo	C1	{1}	{2}	10	10	124.30	120.50	27.34	27.17	0.9924	1.63	0.1210

Las dos pruebas siguientes son simulaciones, obtenida al multiplicar por dos y por ocho la cantidad de casos de las muestras originales, sin modificar las observaciones.

Tabla 2

Prueba t de diferencia de medias, datos ficticios, $n_1=n_2=20$

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)	Var(1)	Var(2)	pHomVar	T	p-valor
grupo	C1	{1}	{2}	20	20	124.30	120.50	25.91	25.74	0.9888	2.36	0.0232

Tabla 3

Prueba t de diferencia de medias, datos ficticios, $n_1=n_2=80$

Clasific	Variable	Grupo 1	Grupo 2	n(1)	n(2)	Media(1)	Media(2)	Var(1)	Var(2)	pHomVar	T	p-valor
grupo	C1	{1}	{2}	80	80	124.30	120.50	24.92	24.76	0.9769	4.82	<0.0001

Puede allí verse, que las medias muestrales son siempre las mismas, como lo es su diferencia, de 3,8. Las varianzas difieren poco y en ningún caso es necesario introducir correcciones ya que a los fines de esta prueba pueden tratarse como homogéneas. Con 10 casos en cada grupo la diferencia no resultó significativa, con 20 en cada uno sí es significativa al 5% pero no al 1% y, cuando se incrementa a 80 casos en cada grupo, el valor p es menor a una diezmilésima, por lo que la conclusión es indubitablemente rechazar H_0, concluyendo que los grupos difieren. Una diferencia de 3,8 puede entonces ser significativa o

no según provenga de muestras de mayor o menor tamaño; es claro que no hay juicio en abstracto acerca de la diferencia, ésta depende de dónde haya sido hallada[5]. Esto sucede porque el efecto del aumento (artificial en nuestro ejemplo) del tamaño de las muestras es el de incrementar el puntaje estandarizado (t), volverlo más extremo, y con ello reducir su probabilidad.

Una respuesta a este problema es calcular medidas de tamaño del efecto. Las medidas de tamaño del efecto tienen en común que eliminan la incidencia de la cantidad de casos, en ese sentido, son medidas estandarizadas, sin embargo, veremos que no están exentas de problemas de interpretación. Se trata de medidas descriptivas que cuantifican la magnitud de una diferencia entre grupos o la intensidad de una relación entre variables. Por eso existen dos grandes familias de coeficientes que evalúan el tamaño del efecto (Ellis, 2010) la de "los d" y la de "los r". La primera familia está compuesta por varios indicadores de la diferencia entre medidas resumen (medias o proporciones) de grupos, que son diferentes según se trate de dos o más de ellos, según los grupos tengan igual o diferente cantidad de casos y según el tipo de variable que se compare. Los más conocidos son, para variables cuantitativas: d de Cohen, Δ de Glass, g de Hedge; y para dicotómicas: riesgo relativo, razón de odds. La segunda familia incluye medidas de intensidad de asociación: r de Pearson, de Spearman, τ de Kendall, y una amplia lista, según el tipo de variable de que se trate.

Para el caso que nos ocupa usaremos como medida el coeficiente d de Cohen, que mide la diferencia entre las medias de dos grupos, expresada en términos de una desviación estándar que se obtiene combinando las de los dos grupos[6]. Dicho de otro modo, indica cuántas desviaciones estándar separan a las medias de los dos grupos.

En las salidas simuladas anteriores este coeficiente es para todos los casos aproximadamente el mismo, en torno a 0,76, lo que indica, según los criterios vigentes, un efecto grande. ¿Cómo interpretar este resultado? ¿Gran efecto, pero no significativo en el primer caso? Hallamos aquí el inconveniente inverso al del valor p, ya que ahora, por no considerar el tamaño de las muestras, dejamos de lado buena parte de la "fuerza de la

[5] En algunos manuales de estadística se advierte sobre el riesgo de hallar resultados estadísticamente significativos cuando las muestras son de gran tamaño, sin embargo esta advertencia suele referirse casi exclusivamente a la prueba chi cuadrado.

[6] $d = \frac{\bar{x}_2 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2}}}$

evidencia".

Además de esta ambigüedad en la lectura del tamaño del efecto, hay un problema adicional: se trata de una medida descriptiva, no inferencial. Puede decirnos lo que sucede a nivel muestral, pero no podemos extrapolar de manera directa ese resultado a toda la población. Los procedimientos para hacerlo no son evidentes y ofrecen estimaciones con errores elevados (Steiger, 2004), luego volveremos sobre este punto.

En la práctica, el tamaño del efecto se usa estableciendo un orden en las operaciones: en primer lugar se dejan de lado los resultados no significativos, de manera de tratar con resultados muestrales que, según la lógica de la prueba de hipótesis, corresponderían a la existencia de algún efecto en la población. Luego, y solo para los resultados que son significativos, se calcula el tamaño del efecto, para evaluar la magnitud de, por ejemplo, una diferencia. Esta forma de proceder pone como criterio determinante la significación estadística.

4. Analizar la potencia

Veamos más de cerca cómo afecta la cantidad de casos al resultado de una NHST: con un aumento en el tamaño de la muestra, manteniendo todo lo demás constante, la prueba tiene más potencia, por lo que es mayor su capacidad para detectar diferencias, aun cuando estas sean pequeñas. La potencia se define como la probabilidad de rechazar una H_0 cuando es falsa, de modo que mide la capacidad de la prueba para detectar diferencias. Es el complemento de la probabilidad de cometer Error de Tipo II (β), por lo que se la abrevia como $1-\beta$. A pesar de la simetría con el Error de Tipo I, hay una diferencia de fondo: si bien el escenario "H_0 verdadera" es único, no sucede lo mismo en el escenario "H_0 es falsa" porque esto es algo que puede suceder de muchas formas. Si la diferencia entre dos grupos no es cero, puede ser cualquier número diferente de cero y todos esos valores corresponden a "H_0 falsa". Por eso no hay un único valor de β (y por consiguiente tampoco de 1-β), como sí lo había con α (0,05; 0,01 u otro valor que nosotros elijamos). Existen tantos valores de 1-β como valores hipotéticos alternativos consideremos. Así, si H_0 es falsa, es porque la diferencia entre las medias poblacionales no es cero, entonces el verdadero valor del parámetro puede ser $\Delta\mu_1$, $\Delta\mu_2$, $\Delta\mu_3$, etc. A cada escenario corresponde una diferente probabilidad de rechazar H_0:

$P(\text{rechazar } H_0/\Delta\mu = \Delta\mu_1)$, $P(\text{rechazar } H_0/\Delta\mu = \Delta\mu_2)$, $P(\text{rechazar } H_0/\Delta\mu = \Delta\mu_3)$

La expresión operacional de "rechazar H_0" es hallar un valor muestral que esté en la zona de rechazo, es decir, más allá de los puntos críticos. Por eso debe calcularse la probabilidad de ese evento (hallar a la diferencia muestral en la zona de rechazo de H_0) bajo cada uno de los escenarios posibles que sean consistentes con "H_0 falsa". El modo más claro de presentar el análisis de la potencia es graficar ese conjunto de probabilidades en función de diferentes alternativas para el parámetro, así se obtiene la *curva de potencia*.

Para las pruebas cuyos resultados se muestran en las salidas 1 a 3, y con nivel de significación del 5%, las curvas de potencia son las siguientes:

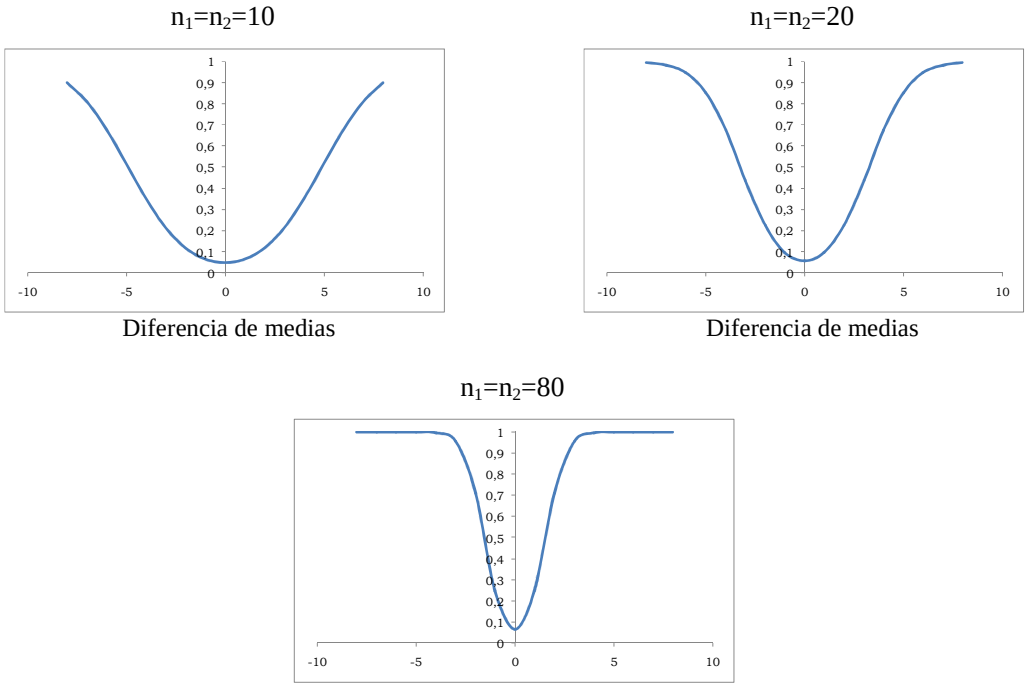


Gráfico 1

Curvas de potencia para la prueba de diferencia de medias con diferentes tamaños de muestra

Fuente: archivo "potencia e IC.xlsx"

Vemos que si se considera como diferencia alternativa entre las medias al valor de la hipotética (cero), la potencia vale en todos los casos 0,05, es decir, el nivel de significación establecido de antemano. Es porque, efectivamente, la probabilidad de cometer un Error de

Tipo I —concluir que las medias difieren cuando no es así—, vale 0,05. El cálculo de la potencia requiere haber fijado el nivel de significación, la capacidad para detectar diferencias siempre lo es a un determinado nivel de significación.

A medida que se consideran valores alternativos más lejanos del hipotético (cero), aumenta la probabilidad de rechazo de H_0. La virtud de estos gráficos es la de ilustrar que la calidad de la prueba se expresa en la "fineza" para detectar diferencias; lo que se refleja en la agudeza de la curva: cuanto más aguda es, tanto más rápido crece la probabilidad de rechazar H_0 al apartarse de cero la diferencia.

El análisis de la potencia es más claro si ésta se expresa en función de alguna medida del tamaño del efecto en lugar de hacerlo en función de valores de la diferencia. Para el ejemplo anterior, calculamos el tamaño del efecto a través del coeficiente d de Cohen y solo nos ocupamos de diferencias positivas (por la simetría de las curvas), podemos así representar la potencia para los diferentes valores de d y distintos tamaños de muestra en un solo gráfico.

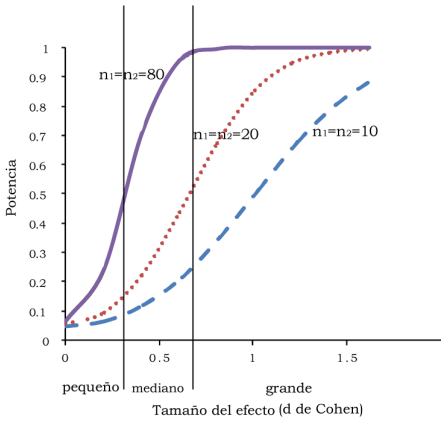


Gráfico 2

Potencia según tamaño del efecto (calculado con d de Cohen) para diferentes tamaños de muestra y significación de 5%

Fuente: archivo "potencia e IC.xlsx"

Puede verse, quizás ahora con más claridad, que la cantidad de casos en la muestra incide sobre la capacidad de la prueba para detectar (rechazando la hipótesis de igualdad de medias) un determinado tamaño del efecto.

Hemos agregado en el gráfico los puntos que delimitan los criterios para considerar de mayor o menor magnitud al tamaño del efecto (Lipsey, 1990), a fin de evaluar con mayor

precisión cómo afecta la cantidad de casos a la probabilidad de detectar diferencias. Con dos muestras de 10 casos cada una, la probabilidad de detectar un efecto mediano es menor al 10%, mientras que si las muestras son de 80 casos cada una, esa probabilidad casi llega al 50%.

Vemos entonces que la dependencia en el tamaño de las muestras es una cualidad intrínseca de las pruebas de hipótesis, su significado es que una misma diferencia hallada en una muestra de mayor tamaño aporta más evidencia para creer que la diferencia no proviene del azar sino de algún efecto incluido en la variable que define los grupos que se comparan. La misma diferencia aporta más evidencia si proviene de una muestra grande que de una pequeña, lo que depende de los recursos dispuestos para la investigación. Su incremento puede llevar a aumentar la potencia de manera artificial, lo que se traduce en la detección de diferencias mínimas, posiblemente sin interés clínico o práctico, pero significativas. Si solo nos interesara hallar resultados significativos y no existiera restricción en los recursos disponibles, siempre sería posible hallar un tamaño de muestra suficientemente grande como para que cualquier diferencia no nula sea significativa, más allá de la magnitud absoluta de esa diferencia. La evaluación del tamaño del efecto ayuda a dilucidar estas situaciones.

En el ejemplo hemos hecho depender la capacidad para detectar efectos solo del tamaño de las muestras, pero está también influida por la dispersión que tenga la variable que se analiza. Esos dos elementos, el tamaño de las muestras y la variabilidad de la característica que se observa son los que determinan si una prueba es capaz de detectar diferencias entre grupos o no. Si la variabilidad de los datos se reduce, la potencia aumenta de manera legítima, porque la esencia del poder explicativo de una variable es que dentro de cada una de sus categorías haya poca dispersión frente a la que hay entre los grupos que definen esas categorías. Pero la reducción de la variabilidad no está bajo control inmediato del investigador, como sí sucede con el tamaño de la muestra. Reducirla implica conocer sus fuentes, identificar las variables que hacen diferencias entre los casos —las que contribuyen a explicar las diferencias—; el problema es contar con una descripción tan completa como sea posible del fenómeno que se observa: cuanto más cabal sea nuestro conocimiento de los factores que inciden en lo que se observa, tantas más fuentes de variabilidad será posible controlar para aislar las variables cuyos efectos nos interesa poner a prueba.

La consecuencia de esta cualidad de las pruebas es la complejización de las comparaciones entre investigaciones. Esta complejización sucede porque no es suficiente

observar si los resultados son significativos o no, sino también considerar el número de casos que se analizan y su variabilidad. En este sentido las pruebas de hipótesis son solo un punto de partida para análisis posteriores.

5. Intervalos de confianza

Los mismos supuestos que sostienen a la prueba de hipótesis permiten calcular intervalos de confianza en torno a un valor muestral hallado para la estimación de medidas descriptivas. Su cálculo es sencillo dado que no requieren otras operaciones que las que conducen a la prueba de hipótesis. Llamamos Θ al parámetro que se estima, $\hat{\theta}$ a su estimador y $\sigma_{\hat{\theta}}$ al error estándar del estimador, entonces los límites del intervalo son:

$$\hat{\theta} \pm t_{1-\alpha/2} * \sigma_{\hat{\theta}}. \qquad [1]$$

Para el caso que nos ocupa se debe contar con la diferencia de las medias muestrales y el error estándar de esa diferencia. Para los ejemplos de arriba, los intervalos, con una confianza del 95%, resultan:

Tabla 3

n_1	n_2	IC 95%	error absoluto	error relativo
10	10	[-1,11; 8,71]	4,91	129%
20	20	[0,55; 7,05]	3,25	86%
80	80	[2,24; 5,36]	1,56	41%

Fuente: archivo “potencia e IC.xlsx”

Los intervalos condensan información acerca del resultado de la prueba de hipótesis y de la magnitud de la diferencia:

- La inclusión o no del cero en el intervalo indica si la hipótesis de igualdad de medias debe aceptarse o rechazarse a un nivel de significación complementario a la confianza usada para construir el intervalo. Que el primero de los intervalos del ejemplo tenga límites de signo diferente —o lo que es lo mismo que incluya al cero—, se corresponde con el no rechazo de H_0 en la prueba de hipótesis de la salida 1.
- La amplitud de intervalo, por su parte, cuantifica la diferencia entre las medias. El error absoluto es la mitad de la amplitud, y nos da información sobre la calidad de la estimación. Se dice que una estimación es más precisa cuando tiene menos error. Una vez fijada la confianza, el error de estimación está determinado por la variabilidad de los datos y el tamaño

de las muestras.

- El error relativo se expresa como porcentaje del estimador puntual, y permite comparar la calidad de estimaciones de magnitud diferente.

Los intervalos de confianza ofrecen más información que el valor p y una idea más clara de los efectos que tratan de estimarse, sea de una medida descriptiva, de una diferencia entre grupos o de un índice de asociación. Además, complementan y amplían el resultado dicotómico de una NHST (de aceptarla o rechazarla). Por eso son recomendados:

> […] sin embargo, un reporte completo de todas las hipótesis puestas a prueba, así como estimaciones adecuadas del tamaño del efecto e intervalos de confianza, constituyen las expectativas mínimas para todas las revistas de APA. (APA, 2010, p. 33, traducción propia)
>
> La inclusión de intervalos de confianza (para estimaciones de parámetros, para funciones de los parámetros, como diferencias de medias y para tamaños del efecto) pueden ser modos extremadamente efectivos para reportar resultados. Debido a que los intervalos de confianza combinan información sobre la ubicación y la precisión y pueden a menudo ser usados directamente para inferir niveles de significación, son, en general, la mejor estrategia de informe. El uso de intervalos de confianza es, por eso, firmemente recomendada. Como regla, es preferible usar un único nivel de confianza, especificado a priori (por ejemplo 95 ó 99%), a los largo de todo el texto. Siempre que sea posible, se debe basar la discusión y la interpretación de resultados en estimaciones puntuales y de intervalo. (APA, 2010, p. 34, traducción propia).

Sin embargo, los intervalos de confianza resultan ser escasamente utilizados; Ellis (2010), citando a varios autores, señala que menos del 2% de los estudios cuantitativos en Psicología los reportan. Cohen (1994) sospecha que su poco uso se explica porque ponen en evidencia la magnitud del error, que suele ser importante: "…they are so embarrassingly large!" (Cohen, 1994, p. 1002).

La mejora en este sentido no debería ser la de ocultar el error, ni hacer crecer los tamaños de muestra para reducirlos, sino buscar el modo de reducir la varianza. Nuevamente aquí se trata de comprender mejor el fenómeno que observamos para conocer fuentes de variabilidad y disminuirla a través del control.

La contrapartida del valor de la información que aportan los intervalos de confianza es que su lectura no es evidente, y son fácilmente malinterpretados. Cuando se determinan los límites de un intervalo para la estimación de un parámetro, se cuenta con dos valores del estimador. La teoría del muestreo afirma que un cierto porcentaje de los intervalos que se

construyan por ese procedimiento, contendrán al parámetro que se estima. Al nivel usual de confianza, un 95% de todos los intervalos que se podrían construir contiene al parámetro. Es decir que nuestra confianza no es en los valores, sino en el procedimiento que usamos para hallarlos (Howell, 2010). Antes de recoger los datos y construir el intervalo, cada uno de los posibles tiene una probabilidad de 95% de contener al parámetro, pero una vez observado el valor del estimador y construido el intervalo, ya no es correcto asignarle una probabilidad, por eso usamos el término *confianza*. La lectura de un intervalo de confianza es que se tiene una confianza $1 - \alpha$ que contenga al parámetro. Un error frecuente en la lectura es afirmar que el intervalo tiene una probabilidad $1 - \alpha$ de contener al parámetro; es un error porque no puede asignar probabilidad a lo que ya se observó. Más equivocada aun es la lectura inversa del intervalo: la probabilidad que tiene el parámetro de estar entre los dos valores ofrecidos. Desde la teoría frecuentista, el parámetro es fijo, por lo que no corresponde asignarle probabilidades. Sin embargo, no es infrecuente encontrar lecturas de este tipo aun en artículos técnicos (Denis, 2003). Este error es el equivalente al de la falacia de la probabilidad inversa en las pruebas de hipótesis, que consiste en creer que el valor p mide de algún modo la probabilidad de H_0. Podríamos parafrasear a Cohen (1994) diciendo que es tan grande el deseo del investigador de reportar entre qué límites se halla eso que quiere estimar, que llega a creer que el intervalo de confianza se lo informa.

El cálculo de los intervalos de confianza se realiza con la expresión [1] en el caso de algunas medidas, pero para otras ―entre ellas muchas de las de tamaño del efecto―, se requieren procedimientos más complejos. Esto se debe a que los intervalos, por ejemplo el correspondiente al coeficiente d de Cohen, no son centrales, los límites no están ubicados simétricos alrededor del estimador y la distribución de probabilidades asociada no es la *t* que conocemos sino una distribución llamada *t no central*. Los intervalos para estimar estas medidas deben calcularse usando métodos numéricos, aproximando sucesivamente, iterando hasta encontrar los límites (Howell, 2010). Estos procedimientos son muy complejos para realizar de manera manual, por lo que se usan métodos computacionales; Geoff Cumming ofrece, en http://www.latrobe.edu.au/psy/esci/, un conjunto de rutinas en hojas de cálculo para construir intervalos de confianza no centrales, en particular la denominada *ESCI DELTA* es de ayuda. En el mismo sitio puede obtenerse *ESCI PPS p intervals*, que permite simular situaciones y comparar las salidas resultantes, es de ayuda para familiarizarse con el uso de intervalos.

Además de la abundante evidencia sobre las limitaciones de las NHST para producir conocimiento científicamente validado, el problema que más nos interesa es la interpretación inadecuada de los resultados de pruebas de hipótesis. Hasta este punto discutimos los modos en que ese problema puede subsanarse, complementando las NHST con otros procedimientos basados en la misma lógica deductiva. En todos ellos sigue vigente la advertencia de Fisher (1959) sobre la imposibilidad de asignar probabilidad a las hipótesis. Esto es así, porque desde este enfoque los elementos que participan de las hipótesis, los parámetros, son fijos; suponemos que tienen un valor determinado, que desconocemos, al que buscamos estimar. Por ser un valor fijo, no corresponde asignarle una probabilidad. En los apartados siguientes nos aproximamos a un enfoque alternativo, para el cual los parámetros son variables y los datos observados, constantes. Será así posible asignar probabilidad a los valores paramétricos, es decir a las hipótesis.

6. El enfoque bayesiano para la estimación

Las dificultades para la interpretación que surgen de la inferencia frecuentista —en forma de pruebas de hipótesis o de intervalos de confianza—, pueden superarse con un cambio de enfoque sobre el problema: el teorema de Bayes aporta una manera diferente de asignar probabilidades y, en consecuencia de realizar estimaciones. Fue enunciado por primera vez por Thomas Bayes (1763) y es una regla que permite invertir el orden de una probabilidad condicional, por lo que, si A y B son dos eventos, sirve para pasar de P(A/B) a P(B/A). El principal interés de este teorema es que permite “aprender de la experiencia”, porque incorpora nueva evidencia para corregir probabilidades ya calculadas. La idea central es la de partir de una probabilidad inicial (a priori) para un evento y luego ajustar su probabilidad a la luz de la evidencia. Se trata de pasar de P(H), que es la probabilidad de una hipótesis asignada antes de contar con las observaciones, a P(H/D), que es la probabilidad de la hipótesis, condicionada a los datos observados; cuando nuevos datos se aportan al experimento, esa probabilidad puede revisarse. Una forma de razonar que resulta familiar ya que, en contextos de incertidumbre, el investigador parte de alguna hipótesis a la que corrige a medida que incorpora información. Ante una demanda de un paciente, el clínico formula hipótesis y a medida que avanza la recolección de datos (entrevista, pruebas) le agrega o quita credibilidad. La hipótesis es revisada tomando en consideración los datos que se aportan, el teorema de Bayes ofrece una técnica para hacer esto de manera formal. Si los eventos que se

consideran son, por un lado el diagnóstico hipotético y por otro la evidencia que se recoge, el teorema de Bayes nos dirá cómo pasar de la probabilidad de "observar lo que se observa si la hipótesis fuera verdadera" a la probabilidad "que tiene la hipótesis de ser verdadera, según lo que se observa", que formalmente implica pasar de $P(datos/H)$ a $P(H/datos)$.

Es un procedimiento que hace un importante aporte al carácter acumulativo de la investigación, del que suele afirmarse que la Psicología adolece (Meehl, 1978). Una crítica hacia la inferencia bayesiana es que requiere partir de una probabilidad (a priori) acerca de la hipótesis, una probabilidad que suele ser subjetiva, por lo que diferentes investigadores podrían asignar diferentes probabilidades a priori y llegar a diferentes resultados usando los mismos datos. Sin embargo, siempre es posible, ya sea fundamentarla o bien establecer una probabilidad a priori "no informativa" que supone que no se dispone de evidencia inicial a favor de ningún valor particular de la hipótesis, aunque la noción de distribución no informativa es discutible (Larsson, 2011).

La formulación del teorema de Bayes adaptada a la inferencia es la siguiente:

$$P(H/datos) = \frac{P(H)*P(datos/H)}{P(datos)} \qquad [2]$$

En la que:

$P(H)$ es la probabilidad a priori asignada a la hipótesis, puede ser subjetiva, provenir de experiencias anteriores, o ser de carácter "no informativo". En este último caso no se asume conocer algo sobre la hipótesis.

$P(datos/H)$ es llamada *verosimilitud*, mide la probabilidad de lo observado si fuera verdadera la hipótesis.

La expresión [2] permite pasar de *P(Datos/H)* que es el resultado de las NHST en el enfoque frecuentista, a la *P(H/Datos)*; que es una información sustancialmente más valiosa, ya que nos indica cuán probable es la hipótesis considerando la evidencia hallada. Sin embargo para llegar allí es necesario hacer algunas puntualizaciones sobre los enfoques, porque el bayesiano difiere del frecuencial en que:

- El parámetro poblacional que se desea estimar (θ) es considerado variable. A diferencia del enfoque frecuentista, desde la teoría bayesiana no se considera al parámetro como un valor fijo, sino como una variable, mientras que lo que se trata como constante es la información que se obtiene en la muestra, es decir los datos observados.

- Para la estimación del parámetro poblacional es necesario un conocimiento previo de la distribución que pueda seguir el parámetro poblacional. Es lo que se llama distribución *a priori*, es una cuantificación de nuestro conocimiento previo acerca del fenómeno. Esto puede provenir de estudios publicados previos o de las expectativas del investigador.
- Cuando se usa para construir intervalos, no se hace referencia a intervalo de confianza, sino a *intervalo de credibilidad*, que es el intervalo que contendría al parámetro poblacional con una probabilidad establecida. Como señalamos antes, es un concepto a veces usado, pero erróneo, para referirse al intervalo de confianza dentro de la estadística frecuentista.

7. Un ejemplo artificialmente simplificado

Para esta presentación introductoria a la inferencia bayesiana, resulta más conveniente comenzar con un ejemplo sobre estimación de la proporción. Supongamos un estudio (Caballero Granado, 2007) en el que se desea conocer la prevalencia de EPOC en una población, para lo que se entrevista a 170 personas. Según la información previa disponible, el investigador considera que lo más probable es que la prevalencia sea del 9%. Un prevalencia del 6% ó del 12% se consideran menos probables, y prevalencias del 2% ó 16% muy improbables. Esta información preliminar constituye una distribución a priori de las probabilidades de los diferentes valores posibles del parámetro. Cuantificamos esta asignación a priori del siguiente modo:

Tabla 1
Probabilidades a priori asignadas a diferentes hipótesis sobre prevalencia de EPOC

Valores hipotéticos de la prevalencia H	Probabilidad a priori P(H)
0,02	0,1
0,06	0,2
0,09	0,4
0,12	0,2
0,16	0,1
Total	1,0

También podría suceder que no se contara con ninguna estimación previa, en ese caso asignaríamos igual probabilidad a los diferentes valores de la prevalencia. Es ese caso se trata de una distribución “no informativa”. Volveremos sobre ese caso más adelante

Cuando se recogen los datos, se encuentra que sobre 170 participantes, 10 confirman la presencia de EPOC, que corresponde a una prevalencia muestral de 5,9%. Desde el análisis bayesiano se observa la probabilidad de este resultado según cada uno de los valores posibles de la prevalencia, y con ello se calcula lo que llamamos la verosimilitud. Debemos responder

a la pregunta ¿Cuán probable sería haber hallado 5,9% en la muestra si el valor poblacional fuera 2%? Y del mismo modo para cada uno de los valores hipotéticos de la prevalencia. Respondemos a eso a través de una distribución binomial:

$$P(\hat{p} = 5{,}9\%/P = 2\%) = B(170;\ 0{,}02,;\ 10) = 0{,}0171$$

Operando del mismo modo, para los demás valores de la prevalencia resulta:

Tabla 6
Verosimilitud de cada valor hipotético a la luz de los resultados observados

Prevalencia: Valores hipotéticos de H	Verosimilitud: P(datos/H)=P(5,9%/H)
0,02	0,00171
0,06	0,12869
0,09	0,04136
0,12	0,00344
0,16	0,00004

Las diferentes hipótesis (H) son los diferentes valores del parámetro y los datos el valor observado en la muestra. Nuestro interés es el de conocer la probabilidad de los diferentes valores de la prevalencia teniendo en cuenta la evidencia observada. Usando el teorema de Bayes, tenemos los siguientes numeradores de la expresión [2]:

Tabla 7
Cálculo del numerador de la fórmula de Bayes

Valores hipotéticos de la prevalencia *H*	Verosimilitud P(datos/H)	Probabilidad a priori P(H)	P(datos/H)*P(H)
0.02	0.00171	0.1	0.000171
0.06	0.12869	0.2	0.025737
0.09	0.04136	0.4	0.016545
0.12	0.00344	0.2	0.000688
0.16	0.00004	0.1	0.000004

La última columna es el producto de las dos anteriores. El denominador de la expresión [2] es fijo, porque es la probabilidad de los datos, y se obtiene como la suma de todas las formas en que podrían resultar estos datos, es decir bajo los diferentes valores posibles de la prevalencia, porque el valor observado puede haber resultado siendo P=2%, 6%, etc. Así entonces:

$$P(datos) = P(\hat{p} = 5{,}9\%) = P(\hat{p} = 5{,}9\%/P = 2\%) * P(P = 2\%) + P(\hat{p} = 5{,}9\%/P = 6\%) * P(P = 6\%) + P(\hat{p} = 5{,}9\%/P = 9\%) * P(P = 9\%) + P(\hat{p} = 5{,}9\%/P = 12\%) * P(P = 12\%) + P(\hat{p} = 5{,}9\%/P = 16\%) * P(P = 16\%)$$

Que es la suma de la última columna de la Tabla 6: 0,04315. Con este valor podemos calcular las probabilidades a posteriori *P(H/datos)* para cada H, usando [2]:

Tabla 8

Cálculo de probabilidades a posteriori

P	*P(datos/P)*	*P(P)*	*P(datos/P)*P(P)*	*P(P/datos)*
0.02	0.00171	0.1	0.000171	0.00397
0.06	0.12869	0.2	0.025737	**0.59652**
0.09	0.04136	0.4	0.016545	**0.38347**
0.12	0.00344	0.2	0.000688	0.01595
0.16	0.00004	0.1	0.000004	0.00008

La última columna indica cuán probable es cada valor de la prevalencia (poblacional) de acuerdo a los datos observados. La lectura del resultado de la primera fila es que, según los datos observados (proporción muestral de 5,9%), la probabilidad que la prevalencia poblacional sea del 2% es de 0,00397. Así pasamos de una distribución de probabilidades asignadas a priori por el investigador en base a su experiencia, a otra distribución, que suma a esa experiencia previa, la información que aportan los datos observados. Esta última es la distribución que usaremos para hacer inferencias.

Vemos que la prevalencia paramétrica más probable es el 6%, que es el más cercano a la proporción muestral. Las dos probabilidades más altas (destacadas en la tabla 7), corresponden a valores de P del 6 y 9%; su suma:

$$0{,}59652 + 0{,}38347 = 0{,}98$$

Se lee como “La probabilidad que tiene la prevalencia poblacional de estar entre 6 y 9% es, según los datos observados y la información previa del 98%”. Este es el que se conoce como *intervalo de credibilidad*. El equivalente de la confianza es ahora la credibilidad y la lectura del intervalo la señala como la probabilidad asociada a los valores del parámetro. Por haber usado un conjunto acotado de valores discretos de la prevalencia no podemos elegir la probabilidad, por ejemplo, al 95%. Eso sí es posible cuando se trata a la prevalencia como variable continua, pero no trataremos ese caso aquí.

Para comparar con el enfoque frecuentista, calculamos el intervalo de confianza de Wald[7] para este ejemplo, con un nivel de confianza igual a la credibilidad de intervalo bayesiano, de 98% (para hacer válida la comparación) y resulta:

$$\hat{p} \pm t_{0,01} * \sqrt{\frac{\hat{p} * (1-\hat{p})}{n}} = 0{,}059 \pm 2{,}35 * \sqrt{\frac{0{,}059 * (1-0{,}059)}{170}} = 0{,}059 \pm 0{,}042$$

Los límites son: [1,7; 10,1] %. La lectura de este intervalo es que "Hay una confianza del 98% que el intervalo [1,7; 10,1] % contenga al valor poblacional de la prevalencia."

Identificamos tres desventajas de este intervalo frecuentista comparado con el bayesiano:

- Su mayor amplitud, indicador de la imprecisión de la inferencia.
- Su dificultad de interpretación, ya que no nos indica la probabilidad de contener al parámetro.
- La falta de uso del conocimiento previo que se tiene sobre el fenómeno que se observa y como consecuencia de esto, la dificultad para acumular experiencia.

Como indicamos antes, una crítica que suele dirigirse contra la estimación bayesiana es que usa probabilidades a priori que son subjetivas. Para remediar esto o bien cuando no se cuenta con información previa, la inferencia bayesiana se realiza asumiendo una distribución a priori "no informativa". Solo trataremos aquí el caso discreto, que es una aproximación, pero que resulta suficiente para dejar planteado el tema.

Nuestra aproximación consistirá en tomar como valores posibles de la prevalencia poblacional, las 99 proporciones contenidas entre cero y uno, desde 0,01 hasta 0,99 y, a diferencia del ejemplo anterior en que el investigador, usando su conocimiento previo, asignó probabilidades diferentes a los valores; supondremos que no hay ninguna razón para creer que la proporción sea una u otra, por lo que trataremos a todas como igualmente probables. La tabla resulta así:

[7] Este es el intervalo de confianza más difundido, sin embargo hay críticas sobre los inconvenientes cuando el resultado de $np(1-p)$ no es lo suficientemente grande (Cepeda-Cuervo et al, 2008). Algunas alternativas a este intervalo son el de Wilson (Newcombe y Merino, 2006) y el de Wald ajustado (Agresti y Coull, 1998). Si se usa la distribución binomial, el intervalo resulta [2,3; 10,6] %, para el cual valen igualmente las comparaciones siguientes.

Tabla 9

Cálculo de probabilidades a posteriori, con una distribución a a priori no informativa

P	*P(datos/P)*	*P(P)*	*P(datos/P)*P(P)*	*P(P/datos)*
0.01	8.4956E-06	0.0101	8.5814E-08	1.4527E-05
0.02	0.00171409	0.0101	1.7314E-05	0.0029311
0.03	0.0191537	0.0101	0.00019347	**0.03275284**
0.04	0.06479775	0.0101	0.00065452	**0.11080418**
0.05	0.1129899	0.0101	0.00114131	**0.19321277**
0.06	0.12868615	0.0101	0.00129986	**0.22005337**
0.07	0.10859627	0.0101	0.00109693	**0.18569966**
0.08	0.07320065	0.0101	0.0007394	**0.12517315**
0.09	0.04136286	0.0101	0.00041781	**0.07073051**
0.10	0.02024717	0.0101	0.00020452	**0.03462266**
0.11	0.00878801	0.0101	8.8768E-05	0.0150275
0.12	0.00344035	0.0101	3.4751E-05	0.005883
0.13	0.00123049	0.0101	1.2429E-05	0.00210413
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
0.98	5.066E-257	0.0101	5.117E-259	8.662E-257
0.99	3.836E-305	0.0101	3.875E-307	6.56E-305
	sumas	1	0.00590702	1

La primera columna es la lista de los valores elegidos de P, hemos tomado 99 entre el 1 y el 99%, pero siempre es una aproximación discreta a los infinitos valores posibles de una variable continua, la usamos con fines expositivos.

En la segunda columna se calcula —a través de una distribución binomial—, la probabilidad del valor observado (5,9%) si la proporción poblacional fuera la que indica la columna 1. Se trata de la verosimilitud. Por ejemplo, la primera probabilidad es la probabilidad de encontrar 5,9% en la muestra si la proporción poblacional fuera de 1%: P(5,9%/1%) y se calcula con una distribución binomial con 170 casos y 10 éxitos (el 5,9%):

$$B(170, 0{,}01, 10) = 0{,}00000849$$

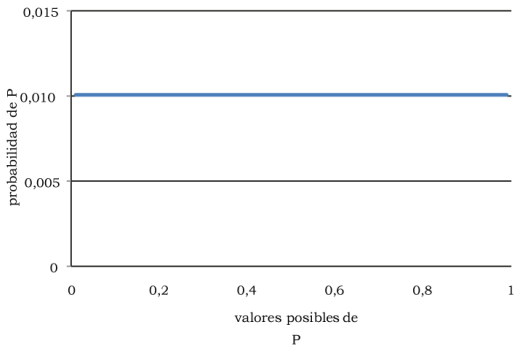
Este último valor de la expresión de arriba se expresa más cómodamente usando notación científica como: 8.4956E-06

La tercera columna es la probabilidad a priori, que asigna a todos los valores de P la misma probabilidad, 1/99 (aproximadamente 0,0101). Aquí se transmite el carácter no informativo de la probabilidad a priori, porque se consideran igualmente probables todos los valores de la proporción paramétrica.

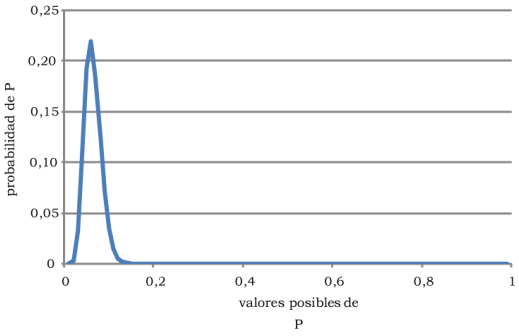
La columna siguiente es el producto de la segunda y la tercera. Su suma provee el denominador de la fórmula de Bayes.

La última columna es el resultado de aplicar la fórmula de Bayes y nos da las probabilidades a posteriori. Como ejemplo de lectura, la primera fila indica que dados los datos que se han observado, la probabilidad que la proporción poblacional sea 0,01 es 1,45E-05.

Esta operación, que consiste en aplicar el teorema de Bayes a una variable a la que estamos tratando como si fuera discreta solo a los fines de la exposición, transforma la distribución de probabilidades a priori —uniforme en este ejemplo—, en la distribución a posteriori, que está expresada en la última columna de la tabla 8. Gráficamente, la distribución a priori es:



Mientras que la distribución a posteriori tiene la forma:



El uso que nos interesa de esta última distribución es el de realizar inferencias, para ello, seleccionamos ahora los valores que concentran el 98% de probabilidad en torno al modo[8] de la distribución (destacados en la tabla 8) y vemos que corresponden a las prevalencias que van del 3 al 10%. Ese es el intervalo de credibilidad para estimar la

[8] Esta no es la única opción posible, ya que una vez que se cuenta con la distribución a posteriori, los intervalos de confianza no son únicos, las formas para definir el intervalo pueden ser:

- Elegir el intervalo más estrecho que delimite la probabilidad establecida como credibilidad. Si la distribución es unimodal, este intervalo contiene al modo. Esta es por la que optamos aquí.
- Elegir un intervalo simétrico en cuanto a la probabilidad, es decir que deje un área por debajo del límite inferior igual a la que queda por encima del superior. Este intervalo contiene a la mediana.
- Centrar el intervalo en la media y alcanzar la credibilidad establecida de manera simétrica a su alrededor.

prevalencia poblacional. La lectura del intervalo es que “Hay una probabilidad del 98% que la prevalencia en la población esté entre el 3 y el 10%”.

Como vemos, al no usar la información a priori obtenemos un intervalo más amplio que antes, con mayor error de estimación. Esto ilustra el valor que tiene el uso de información previa para mejorar la precisión de la estimación, es decir para reducir la amplitud del intervalo. No obstante el error obtenido en este segundo ejemplo sigue siendo menor que el que presentaba el intervalo frecuentista.

Resumimos los resultados de los ejemplos presentados. Con una muestra de 170 casos en la que se encuentran 10 casos positivos y al 98%, tenemos:

- Intervalo de confianza (enfoque frecuentista): [1,7; 10,1] % (ó [2,3: 10,6] % si se usa modelo binomial)
- Intervalo de credibilidad (enfoque bayesiano sin información previa): [3; 10] %
- Intervalo de credibilidad (enfoque bayesiano con información previa): [6; 9] %

La mejor calidad de la estimación se logra cuando se integra la información a priori con lo que se observa, ya que en ese caso se suma lo que se conoce con anterioridad, para mejorar la estimación a la luz de nuevos datos. Pero aun en el caso que se acepte la crítica a la estimación bayesiana y se evite el uso del conocimiento previo disponible, el segundo el intervalo tiene más precisión que el frecuencial y agrega la ventaja de una lectura más intuitiva.

8. Conclusión

Como hemos mostrado, los procedimientos estadísticos no están exentos de contradicciones y limitaciones, ninguno es definitivo ni reemplaza a la reflexión del investigador. Mientras el enfoque frecuentista siga siendo hegemónico, debe tenerse en cuenta que el mero reporte de la significación estadística es insuficiente y debe acompañarse de información sobre la magnitud de los efectos hallados, así como de los riesgos de cometer error de tipo II (y su complemento, la potencia). Aunque ninguna de esas medidas sea suficiente por sí misma, en conjunto aportan para ofrecer una mejor perspectiva de interpretación de resultados de las pruebas de hipótesis.

Por su parte, los intervalos de confianza condensan un gran volumen de información, que los hace convenientes en la comunicación de resultados. Para su construcción, al estimar la proporción, existen opciones: el cálculo exacto de los límites con distribución binomial, la aproximación de Wald y algunas alternativas que mejoran su calidad. Construir intervalos para medidas más complejas, como las de tamaño del efecto, requiere procedimientos de iteración numérica, pero hay rutinas disponibles para hacerlo de manera automática.

El enfoque bayesiano tiene la ventaja de dar como resultado probabilidades asociadas

a diferentes hipótesis a partir de los datos observados, pero requiere un cambio muy marcado en la cultura del análisis estadístico, porque trata a los parámetros como variables, no como valores fijos, que es un postulado básico del enfoque frecuencial.

En la construcción de intervalos de credibilidad, se admite optar o no por el uso de información previa, por lo que son flexibles y se adaptan a etapas avanzadas de investigaciones que han acumulado conocimiento que puede usarse para mejorar las estimaciones, o bien a etapas iniciales cuando poco se conoce del tema que se aborda.

Como disciplina viva, la Estadística está en continua revisión metodológica y también de las condiciones de aplicación de los modos de hacer que propone; la creencia en un conjunto de reglas técnicas inmutables no se sostiene, hay en este tema tanta discusión como en cualquier campo de conocimiento donde se confrontan ideas. Quizás la enseñanza tradicional de la Estadística haya traslucido entre los estudiantes que se inician y también en investigadores que son usuarios de la Estadística, una concepción estática, casi dogmática, de contenidos terminados y completos para su uso, procedimientos y reglas de aplicación no cuestionada. Las tensiones y las discusiones que pueden leerse en publicaciones, muestran lo equivocado de esa concepción imaginariamente pura y sin grietas. Por el contrario, la actual es una época en que la necesidad de cambios se está expresando de manera muy amplia, las críticas y la existencia de alternativas accesibles, hacen suponer que habrá modificaciones profundas en los próximos años, el interés de la APA por analizar en detalle esas críticas y su traducción en los ajustes que aparecen en la edición 2010, confirman que ya no son investigadores aislados los que solicitan cambios de mirada en el uso de procedimientos estadísticos.

Los usuarios debemos estar a la altura de la dinámica de la disciplina, porque es peligroso tanto aplicar mecánicamente la más compleja y moderna de las técnicas que aparezcan en la literatura, como aferrarnos a las conocidas porque se han mostrado eficaces; nuestra principal necesidad es la de conocer sobre qué base lógica operan los procedimientos. Informarnos sobre las discusiones, participar en los debates sobre las posibilidades y restricciones de las técnicas que usamos y sus modos de razonar, nos da flexibilidad para su aplicación. Un uso verdaderamente herramental de la Estadística requiere competencia para dar fundamento a una elección al desestimar otras opciones. Eso facilita la apropiación crítica de nuevos procedimientos y, en especial, amplía el horizonte de lectura e interpretación de los resultados. La transformación de las salidas informáticas en texto significante se enriquece si conocemos el fondo del procedimiento que usamos, si podemos dar cuenta de todas la implicaciones del resultado numérico que genera cualquier paquete estadístico.

Referencias

Agresti, A. & Coull, B. (1998). Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician* vol. 52, nº 2, 119–126.

APA (2010). *Publication Manual of the American Psychological Association* Sexta Edición. APA: Washington, DC.

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London* 53: pp. 370–418.

Bellhouse, D. (2004). The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth, *Statistical Science*. Vol. 19, No. 1, 3–43. Disponible en http://www.york.ac.uk/depts/maths/histstat/bayesbiog.pdf

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*. No 33, 526-542.

Buckingham, J., Smith LeBeau, L. & Klein, W. (2011). The performance versus ability distinction following social comparison feedback. *Current Research in Social Psychology*. Disponible en http://www.uiowa.edu/~grpproc/crisp/crisp16_7.pdf accedida el 20/4/2011

Caballero Granado, F. (2007). *Inferencia estadística según el modelo bayesiano* Sociedad Andaluza de Enfermedades Infecciosas. Disponible en http://saei.org/hemero/epidemiol/nota3.html accedida el 5/3/2011

Cepeda-Cuervo, E., Aguilar, W., Cervantes, V., Corrales, M., Díaz, I. & Rodríguez, D. (2008). Intervalos de confianza e intervalos de credibilidad para una proporción. *Revista Colombiana de Estadística* vol. 31, nº 2, pp. 211 a 228. Disponible en http://www.kurims.kyoto-u.ac.jp/EMIS/journals/RCE/V31/bodyv31n2/v31n2a06CepedaEtAl.pdf accedida el 5/3/2011

Coello, M. y Fernández, J. (2011). Actitudes hacia las mujeres de los esquemáticos frente a los no esquemáticos de género *Psicothema* Vol, 23, nº 2, pp, 180-188 Universidad Complutense de Madrid, Disponible en http://www.psicothema.es/pdf/3868.pdf accedida el 20/4/2011

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

Cumming, G. (2011). Exploratory Software for Confidence Intervals, ESCI http://www.latrobe.edu.au/psy/esci/ accedida el 26/6/2011

Denis, D. (2003). Alternatives to null hypothesis significance testing. *Theory & Science* Vol 4

Díaz Batanero, C. (2007). *Introducción a la Inferencia Bayesiana*. Granada: La autora.

Díaz Batanero, C. (2007). *Viabilidad de la Enseñanza de la Inferencia Bayesiana en el Análisis de Datos en Psicología*. Tesis doctoral. Universidad de Granada. Departamento de Psicología Social y Metodología de la Ciencias del Comportamiento. Disponible en http://hera.ugr.es/tesisugr/16582664.pdf accedida 10/12/2010

Díaz, C., Batanero, C. & Wilhelmi, M. (2008). Errores Frecuentes en el Análisis de Datos en Educación y Psicología. *Publicaciones.*

Eberly, L. & Casella, G. (2003). Estimating Bayesian credible intervals. *Journal of Statistical Planning and Inference* 112 115 – 132. Disponible en www.elsevier.com/locate/jspi accedida el 24/05/2012

Ellis, P. (2010). *The Essential Guide to Effect Sizes Statistical Power, Meta-Analysis and the Interpretation of Research Results*. Cambridge University Press:UK

Fisher, R.A. (1959). *Statistical methods and scientific inference*. New York: Hafner Publishing.

Haller, H. & Krauss, S. (2002). Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research Online 2002*, Vol,7, No,1 Institute for Science Education. Disponible en http://www.dgps.de/fachgruppen/methoden/mpr-online/issue16/art1/haller.pdf accedida 5/10/2010

Howell, D. (2010). Confidence Intervals on Effect Size. Disponible en http://www.uvm.edu/~dhowell/methods8/Supplements/Confidence%20Intervals%20on%20Effect%20Size.pdf accedida 3/7/2011

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement.* 56(5), 746-759.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*. Vol 56(1) 16-26.

Larsson, R. (2011). How Informative is a Noninformative Prior? *Documento de Trabajo 2*. Departamento de Estadística. Universidad de Upsala. Disponible en www.statistics.uu.se accedida el 30/5/2012

Lipsey, M.W. (1990). *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, CA: Sage.

Meehl, P. (1967). Theory-testing in Psychology and Physics: A methodological paradox *Philosophy of Science*. Vol, 34, 103–115

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,

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. En L. L. Harlow, S. A. Mulaik, & J.H. Steiger (Eds.) *What if there were no significance tests?* (pp. 393-425). Mahwah, NJ: Erlbaum.

Newcombe, R. & Merino, C. (2006). Intervalos de confianza para las estimaciones de proporciones y las diferencias entre ellas. *Interdisciplinaria* 23, 141–154.

Orlitzky, M. (2012). How Can Significance Tests Be Deinstitutionalized? *Organizational Research Methods* vol. 15, no. 2, 199-228

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods.* 1(2), 115-129.

Shrout, P. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, 8, 1-2.

Steiger, J. H. (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(2), 164-82.