Epidemiología Clínica

Revisando la literatura médica.
Significancia estadística: $p<0,05$

APPRAISING MEDICAL LITERATURE.
Statistical significance: $p<0,05$

JUAN PABLO CHALCO ORREGO*

Desde su implementación la Teoría de la evaluación de significancia de la hipótesis nula (valor del *p*) ha sido criticada en su metodología así como en su significado. Sin embargo es el método más usado en los reportes de trabajos que se publican en medicina y otras ciencias como intento de evaluación de un fenómeno. El test de nivel fijo consiste en proponer una hipótesis nula (no diferencia entre grupos para una resultante) con un nivel critico o nivel de significancia (usualmente 5% o 0,05) en un afán de encontrar un resultado extremo (*p* significativo < 0,05) que rechace la hipótesis nula. Se llama resultado extremo porque se asume en el fondo la igualdad entre resultados o características y el tener un *p* significativo o crítico era un resultado " poco probable " al comparar la gama de posibilidades en una distribución normal de la diferencia encontrada (¿Es la diferencia encontrada debida al azar?). Ver Figura 1.

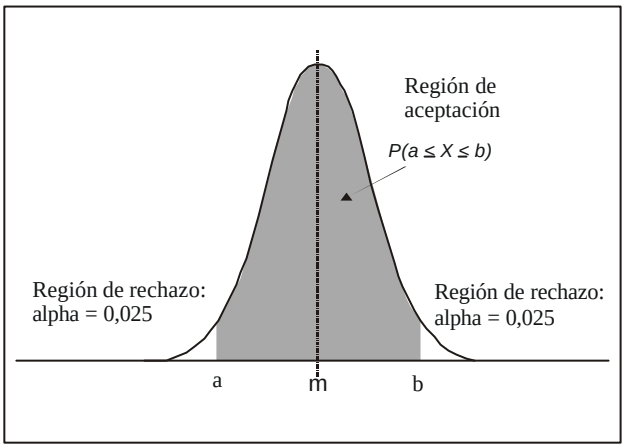

Figura 1.- Distribución normal asumida de la diferencia de resultados encontrados al comparar dos grupos. Prueba a dos colas.

El problema de la validez del *p* empieza con la elección de una **buena muestra** (tamaño adecuado) pues no debe buscarse *p* significativo si la muestra no es lo suficientemente grande para poder hallarla.

La **elección de la prueba** es otro punto importante para la validez de la significancia. Existen más de 6 maneras de calcular el *p* para variables categóricas (ordinales o nominales) y numéricas. La elección está basada en: a) En si los resultados son categóricos o numéricos. b) Si las muestras son pareadas o no. Un ejemplo para entender este punto es al querer medir el peso promedio entre escolares de sexo masculino y femenino. Si juntamos todos sus pesos y sacamos sus respectivas medias esta es una prueba *no pareada***.** Pero si por cada niño que pesamos con una talla determinada busco una niña de la misma talla y también la pesamos esta es una muestra *pareada.* c) Por ultimo nos interesa saber si la muestra es grande o pequeña para lo cual la mayoría toma el límite de 40 por grupo (más de 40 muestra grande y viceversa) para las variables nominales. Y también influye, para variables ordinales o numéricas, si son dos o más de dos los grupos a comparar al evaluar la diferencia significativa entre ellos. Para cada una de estas opciones existe una prueba precisa para calcular el *p*. Ver Tabla 1.

* Médico Pediatra Asistente IESN.

Tabla 1.- Elección de la prueba.

Tipo de variable	Grupos no pareados	Grupos pareados
Nominal		
Muestra pequeña	Test exacto de Fisher	Test Sign
Muestra grande	Chi cuadrado	Chi cuadrado de Mac Nemar
Ordinal		
Dos grupos	Test para dos grupos de Wilcoxon o Test de Mann Whitney U	Test rankeado de Wilcoxon
Más de dos grupos	Análisis de varianza a una vía de Kruskal Wallis	Análisis de varianza a dos vías de Friedman
Numérico		
Dos grupos	T de Student	T de Student pareado
Más de dos grupos	Test F	

Es obvio que las formulas y los resultados difieren si se elige mal o inadecuadamente estas pruebas. Para mayor detalle de cada una de ellas favor remitirse a un libro de Estadística Médica.

Por otro lado **si la significancia es cierta** esta *p*uede ser **malentendida** por factores de confusión no buscados o analizados que alteren el resultado significativo del *p*. Tomemos el ejemplo de un trabajo para evaluar la diferencia de IQ entre el orden de nacimiento de los hijos donde incluiríamos a todas los hijos de familias con un solo hijo, dos hijos, tres hijos, etc. ¿qué pasa si encontramos que el IQ disminuye significativamente a medida si es el primer hijo, segundo o tercero? ¿Debemos alertar a los los padres que no tengan más de un hijo? O sencillamente al tomar las muestras todos las familias con un solo hijo (que aparecían como primogénitos) probablemente tengan mayor cultura para no tener tantos hijos y den mayor estímulo a su único hijo que el segundo o tercero hijo de una familia mayor con menos cultura. En este caso no influye el orden de nacimiento sino el número de hijos de cada familia incluida que refleja la cultura y el estímulo que dan a sus hijos. Se debió incluir solo familias numerosas para comparar si influye este factor.

Para finalizar existe una diferencia entre significancia estadística y clínica que trataremos de explicar con este ejemplo sencillo.

Tomemos estos seis trabajos (tabla 2) que buscan encontrar diferencias significativas entre niveles séricos de colesterol entre dos grupos de intervención. Se muestran las diferencias de medias de los valores de colesterol (x_1-x_2), la desviación estándar de la diferencia (SE), el valor t de la diferencia (t de student), el *p* calculado (por t de student), si el *p* fue significativo, el intervalo de confianza y finalmente su relevancia clínica. Es obvio que para todo este calculo se asumió la normalidad de la curva de la diferencia de medias para el cálculo del *p*.

El trabajo 1 por ejemplo tiene un *p* significativo pero el intervalo de confianza de la diferencia es pequeño muy cercano a cero (no diferencia) por lo tanto no tiene significancia o relevancia clínica. El trabajo 2 además de tener un *p* significativo tiene el intervalo de confianza bastante lejos de cero y es el único que tiene significancia clínica. El trabajo 3 tiene un *p* significativo pero su intervalo de confianza se acerca en su valor mínimo al cero pero a diferencia del primer trabajo no se puede afirmar solo de esta manera si es relevante o no. Los

Tabla 2.- Cálculo de la significancia estadística y apreciación de su significancia clínica.

Trabajo	$\overline{X}_1 - \overline{X}_2$	SE	t	p	p<0,05	95% IC	Relevancia Clínica
1	2	0,5	4	<0,0001	Si	(1,3)	No
2	30	5	6	<0,0001	Si	(20,40)	Si
3	30	14	2,1	0,032	Si	(2,58)	?
4	1	1	1	0,317	No	(-1,3)	No
5	2	30	0,1	0,947	No	(-58,62)	?
6	30	16	1,9	0,061	No	(-2,62)	?

demás trabajos no tienen significancia estadística y mucho menos clínica. En el próximo número tocaremos el tema de cómo evaluar los resultados clínicamente significativos.

Tortura los datos lo suficiente y te confesará cualquier cosa que quieras.

Bulwer-Lytton (1803-1873)

Si calculas con seis pruebas distintas el p y en solo una de ellas te sale significativo.....¿Con cuál resultado lo publicas?

Anónimo

Una teoría sólo tiene dos posibilidades: o es cierta o es falsa; un modelo tiene una tercera: puede ser cierta pero irrelevante.

Manfred Eigen (1927-)

REFERENCIAS BIBLIOGRÁFICAS

1. Bruno Lecoutre. Beyond the significance test controversy: Prime time for Bayes? Leído el 12/04-04 http://www.stat.fi/isi99/proceedings/arkisto/varasto/leco0735.pdf
2. Varkevisser CM. Designing and conducting health systems research projects: volume 2 Leído el 12/04-04 http://web.idrc.ca/en/ev-33013-201-1-DO_TOPIC.html
3. Nielsen J. Probability Theory and Fishing for Significance. Alertbox, March 1, 2004 Leído el 12-04-04 http://www.useit.com/alertbox/20040301.html
4. Sacket DL, Haynes RB, Guyat GH.Clinical Epidemiology: a basic science for Clinical Medicine. Little Brown:1991.
5. D.L. Sacket DL, Richardson WS, Rosemberg W. Evidence Based Medicine: How to Practice and Teach EBM. Churchill Livingstone: 1997.
6. Mould R. Introductory Medical Statics. Publishing Bristol & Philadelphia:1998.
7. Petrie A, Sabin C. Medical statics at a glance. Blackwell Science Ltd. 2000.

Correspondencia:
Dr. Juan Pablo Chalco Orrego
E-mail: jpcho@ec-red.com