Cómo leer un artículo:
Tamaño muestral y poder estadístico: ¿Para que sirven?

Sample size and statistic power: What are they useful for?

Laura Ninín[‡], Gabriel Villalón[‡], Fernando Rubinstein[‡] y Sergio Terrasa[‡].

Resumen

Este artículo describe algunos conceptos referidos a la selección de las muestras para las investigaciones clínicas y los dos tipos de errores que pueden afectar la interpretación de sus hallazgos: el error de tipo I ó a en los estudios cuyo resultado ha sido "positivo" y el error de tipo II o ß, debido a una falta de poder estadístico en los que el resultado ha sido"negativo". Con ejemplos clínicos se describe la importancia de este último en los estudios de equivalencia.

Palabras clave: error aleatorio, bioequivalencia, poder estadístico, tamaño muestral.

Ninín L, Villalón G, Terrasa S. Tamaño muestral y poder estadístico: ¿para qué sirven? Evid. actual. práct. ambul; 10(5): 148-150, Sept-Oct.2007.

Introducción

Cuando leemos artículos de investigación clínica es muy común que nos encontremos con una parte de la sección "métodos" en la que se hace referencia a conceptos como "error alfa" (a), valor de la "p", tamaño muestral, "error beta" (ß) y/o poder estadístico.
El objetivo de este artículo es repasar su significado y discutir la información que estos conceptos pueden aportarnos.

El Muestreo

Todo estudio de investigación nace con una pregunta específica, una de cuyas caracterìsticas más salientes es que debe ser posible de responder, es decir; ¿cuán factible es realizar un estudio que pueda responder adecuadamente esta pregunta? Para comprender esto, debemos primero saber que al momento de diseñar un estudio de investigación resulta esencial estimar el número de observaciones o participantes necesario para responder la pregunta original (tamaño de la muestra con la que se trabajará). Salvo en las excepciones de los censos, casi nunca podemos trabajar con el universo entero de la población a describir o analizar -debido a que resultaría costosísimo en términos de tiempo y recursos. Los investigadores seleccionan una muestra que debería ser representativa de dicho universo, para poder trasladar los hallazgos obtenidos al final de la investigación, a la población destinataria.

Podemos decir que para que una muestra sea representativa de un universo al que se intenta trasladar sus conclusiones, todos los integrantes de dicho universo o población deberían tener la misma probabilidad de haber sido seleccionados para integrar dicha muestra. Por ejemplo, será más representativa una muestra seleccionada en forma aleatoria sobre el padrón de todos los integrantes de una comunidad que una selección "por conveniencia" de los que tienen abierta su historia clínica en el centro de salud, debido a que en este último caso podría objetarse que los que han concurrido al centro de salud pueden ser algo diferentes a los que no lo han hecho y, eventualmente, no representativos del total de la comunidad.

En cuanto al tamaño de la muestra y los errores estadísticos resultantes de una mala planificación de la investigación pueden clasificarse básicamente en dos tipos de problemas:
1) Que el tamaño de la muestra resulte insuficiente y que el estudio no haya podido identificar o demostrar una asociación estadística que realmente existe entre el evento de interés y algunas de las características de los grupos analizados.
2) Que el tamaño de la muestra sea demasiado grande respecto del necesario para identificar la relación y esto nos lleve a considerar la existencia de diferencias "estadísticas" cuando no existen desde el punto de vista biológico o epidemiológico.

A veces es útil pensar en el tamaño muestral como la lente con la que se buscan las diferencias entre los grupos. Una lente de escaso aumento puede limitarnos la posibilidad de ver las diferencias o efectos que realmente existen (escaso poder) y una con excesivo aumento nos magnificará diferencias en detalles a veces insignificantes desde el punto de vista clínico o epidemiológico.

Por lo tanto, la estimación de la cantidad de observaciones (o participantes) necesarios para, identificar o demostrar una diferencia o un efecto en los resultados de un determinado tratamiento (comparado con otro tratamiento o placebo) o la asociación entre dos o más características (p.ej. la asociación entre determinado hábito de vida y el riesgo de padecer cierta enfermedad) constituye la determinación del tamaño muestral.

Es el investigador quien debe preocuparse por estimar correctamente el aumento que su lente necesita en función de lo que quiere observar y existen fórmulas específicas para determinar correctamente el tamaño muestral para cada tipo de estudio que exceden el objetivo de esta nota. Sin embargo, es útil saber que depende fundamentalmente de tres características: 1) el azar, 2) la magnitud de la asociación o el efecto que se espera observar y 3) el poder del estudio para identificar esa diferencia.

Quien, como la mayoría de nosotros, necesita interpretar los resultados, solo debe saber donde mirar para reconocer rápidamente si el tamaño del estudio es el adecuado para responder la pregunta y esto esta expresado tanto gráfica como numéricamente en los llamados "intervalos de confianza", sobre los que escribiremos en la próxima entrega.

Las hipótesis

En la mayoría de los estudios analíticos de investigación cuantitativa existe una hipótesis a demostrar o refutar. Por ejemplo, que un tratamiento tiene mejores resultados que otro tratamiento (o la ausencia de tratamiento); que un tratamiento nuevo más corto, más barato o con menores efectos adversos es por lo menos igualmente efectivo que el estándar de tratamiento actual; que una prueba diagnóstica tiene mejor o igual desempeño que otra considerada como la prueba estándar o de referencia para ese momento histórico; que existe una determinada asociación o no entre dos o más características o variables (por ejemplo entre el hábito de fumar y la probabilidad de desarrollar enfermedad cardiovascular); que una población tiene determinadas cualidades en relación a otra, etc.

Cuando la hipótesis que se pone a prueba es la de que no existen diferencias entre los grupos o tratamientos se habla de una hipótesis nula. Por el contrario, cuando la hipótesis que se con-

* Servicio de Medicina Familiar y Comunitaria del Hospital Italiano de Buenos Aires. sergio.terrasa@hospitalitaliano.org.ar

trasta afirma que existe una diferencia real entre ambos grupos o tratamientos, se habla de una hipótesis alterativa.

Por otro lado, el resultado de un estudio de investigación puede estar reflejando la realidad (la "verdad del universo") o bien, brindar información distorsionada que no corresponde a dicha realidad. Cuando esto último ocurre, puede ser debido a varias razones que pueden ser clasificadas básicamente en tres grandes subgrupos: 1) el efecto del azar; 2) los sesgos o errores sistemáticos del diseño del estudio; 3) los factores de confusión ("confusores" o "confundidores").
Nos referiremos en este artículo a los errores que dependen del efecto del azar o errores aleatorios.

Errores debidos Al azar

Estudios que intentan probar que existen asociaciones o diferencias

Si como resultado de un estudio realizado sobre una muestra representativa del universo observamos que existe alguna asociación estadística (por ejemplo, diferencias en el patrón alimentario entre dos subpoblaciones, mejores resultados en los pacientes que fueron sometidos a un tratamiento respecto de los que recibieron placebo, etc.) estaríamos rechazando en nuestro análisis a la hipótesis nula. Recordamos que en este caso, la hipótesis nula es la que sostiene que no hay asociación estadísticamente significativa entre ambos grupos (en el primer ejemplo, que ambas subpoblaciones se alimentan de la misma forma y en el segundo, que recibir la droga activa es similar a recibir placebo).
Sin embargo, puede ocurrir que el hallazgo de nuestro estudio (el rechazo de la hipótesis nula) no se corresponda con la "verdad" del universo, sino que la diferencia entre grupos o el efecto que observamos haya ocurrido solo a consecuencia del azar.

Lamentablemente, es imposible eliminar completamente el error aleatorio (de tipo I ó a). Lo único que pueden hacer quienes diseñaron el trabajo es estimarlo y comunicar al lector la probabilidad de que este haya ocurrido. De acuerdo a los valores de cada comunidad, a la temática que se está investigando, a la decisión que pueda surgir de dicha información, etc., dependerá el grado de incertidumbre que se tolerará respecto de los hallazgos de la investigación.

Para estimar la probabilidad de error de tipo I se utilizan pruebas de contraste de hipótesis o de significación estadística. Estas pruebas estiman qué probabilidad hay de que la asociación encontrada en la investigación (por ejemplo, la diferencia entren los resultados de dos grupos sometidos a distintos tratamientos) haya ocurrido por azar. Como dijimos, sus resultados expresan una probabilidad (p) que se conoce como significación estadística o valor de la p. Basándonos en esta probabilidad decidiremos o no, rechazar la hipótesis nula. Cuanto menor sea esta probabilidad (cuanto menor sea el valor de la p) mayor evidencia habrá en contra de la hipótesis nula. En este caso se habla de estudios con resultados positivos o que encuentran diferencias.

Para la mayoría de los estudios de investigación se acepta como máximo una probabilidad de azar de hasta un 5% o bien, una posibilidad en 20 de que los hallazgos de la investigación hayan ocurrido por errores puramente aleatorios. Otro modo un poco más complejo de expresar este concepto es el siguiente: si en el universo al que intentaremos extrapolar nuestros hallazgos no existiera la asociación que estamos investigando (si la hipótesis nula fuera la correcta) y si repitiéramos 20 veces dicho experimento (nuestra investigación) sólo en una de esas 20 rechazaríamos en forma errónea la hipótesis nula.

Para ilustrar al lector de donde surge ese 5% tan consensuado y repetido, pensemos en la probabilidad de que una moneda caiga del mismo lado entre cuatro y cinco veces seguidas. Esta situación, si bien existe la probabilidad de que ocurra, es muy poco probable que sea debida solo al azar.

Estudios que intentan probar que no existen asociaciones o diferencias

Cuando los resultados de un estudio muestran que no hay asociación entre dos variables o que los resultados de dos estrategias terapéuticas son equivalentes, los llamamos estudios con resultados negativos o, coloquialmente, estudios negativos.
Sin embargo, puede ocurrir que aún existiendo en la realidad tal asociación, por ejemplo una diferencia clínicamente relevante entre dos tratamientos, el estudio haya sido incapaz de detectarla como estadísticamente significativa. Si esto ocurriera, concluiríamos erróneamente que no existen diferencias o efectos cuando en realidad estas existen. Este error se conoce como error de tipo II ó ß , o insuficiente poder estadístico y es en general atribuible a un número de participantes o a un número de características o eventos menores del estimado inicialmente.

La probabilidad de cometer un error depende del poder la investigación. Cuanto más poder para detectar diferencias o asociaciones estadísticas tenga el experimento (por ejemplo el ensayo clínico) menor será la probabilidad de que se cometa error de tipo II o ß. Por eso, al complementario del error tipo II ó ß (1- ß) se lo conoce como poder estadístico o potencia estadística.
El poder estadístico representa la probabilidad de identificar asociaciones entre variables cuando estas realmente existen. Dicho de otra manera, representa la capacidad o sensibilidad de una prueba (en este caso nuestra investigación en cuestión) para detectar como estadísticamente significativas diferencias o asociaciones de una magnitud determinada.

Por eso es importante recalcar que cuando un estudio arroja un resultado negativo y especialmente si ese resultado negativo puede implicar un cambio de conducta, será muy importante verificar el poder que ese estudio tenía para detectar la diferencia que había estimado entre los grupos, determinada, en general por el sentido común desde lo que considera la comunidad científica como clínicamente relevante.

Por ejemplo, si el estándar actual de tratamiento con penicilina de la faringitis estreptocócica es de diez días de duración y queremos saber si un tratamiento más corto -por ejemplo tres días- tiene la misma efectividad que el estándar, va ser muy importante no cometer error tipo II o ß. Nos va a importar mucho la magnitud de la diferencia que queremos que no se nos "escape" poder documentar y tenemos que asegurarnos que esté explicitada en la sección de materiales y métodos de la investigación. En este caso, podría tratarse de un 10% de diferencia en la tasa de erradicación del Estreptococo beta hemolítico del grupo A (EBHA) medida a los siete días de concluido el tratamiento completo.

Dependerá de nuestro juicio clínico si nos parece adecuada o no dicha magnitud del efecto.

A este tipo de estudio se lo denomina de bioequivalencia y suele exigírsele un poder alto, de un mínimo de 0,9 (90%) ó 0,95 (90%). En este caso queremos que el error tipo II o ß no supere el 0,1 (10%) o el 0,05 (5%) ya que una conclusión errónea implicaría que las personas comenzaran a ser tratadas con sólo tres días de penicilina, cuando quizás no tenga realmente la misma efectividad. Imaginemos la multiplicación poblacional de una decisión tomada por asumir en forma errónea que dos tratamientos son igualmente efectivos cuando realmente no lo son.

Vale aclarar que cuanto menor sea la diferencia de la magnitud del evento que queremos detectar, necesitaremos mayor cantidad de observaciones para poder hacerlo (una lente de mayor aumento). Volviendo al ejemplo anterior, será necesario contar con más pacientes por grupo de tratamiento para detectar una diferencia de un 10% en la tasa de erradicación del EPHA que para detectar una diferencia de un 30%.

A modo de conclusión, la tabla 1 resume conceptualmente los dos tipos de errores aleatorios que puede cometer una investigación.

Tabla 1: tipos de errores aleatorios.

		Realidad (la verdad del Universo)	
		Existe la asociación o diferencias entre los grupos	**No existe la asociación o diferencias entre los grupos**
Resultados de la investigación	**Encuentra la asociación o** diferencias entre los grupos	Resultado verdadero positivo	Resultado falso positivo o error tipo I α
	No encuentra la asociación o diferencias entre los grupos	Resultado falso negativo o error tipo II o β (falta de poder)	Resultado verdadero / negativo

Bibliografía Recomendada

Pértegas Díaz S y Pita Fernández S. Cálculo del poder estadístico de un estudio. Cad Aten Primaria 2003; 10: 59-63. Disponible en URL: http://www.fisterra.com/mbe/investiga/poder_estadistico/poder_estadistico2.pdf (último acceso 17/09/07).
Rubinstein F. Medicina Basada en la Evidencia. En: Rubinstein, A, Terrasa S, Durante E, Rubinstein E, Carrete P, Zárate M y Barani M editores. Medicina Familiar y Práctica Ambulatoria. Buenos Aires: Editorial Médica Panamericana; 2006. Capítulo 7. p. 64-85.
Terrasa S, y col. El reporte de un caso y las series de casos. Evid. actual. práct. ambul; 10(1): 19-22, ene-feb.2007.
Disponible en URL: http://www.foroaps.org/files/guia%20de%20serie%20de%20casos.pdf (último acceso 18/11/07).

Recibido el 20/08/07 y aceptado el 18/09/07.

Cómo leer un Artículo

De un vistazo…

Estrés crónico en el trabajo y el síndrome metabólico: estudio prospectivo.

Chandola T., Brunner E., Marmot M. BMJ 2006;332:521-5.

El objetivo de este estudio prospectivo de cohortes* es investigar la asociación entre el estrés en el trabajo y el síndrome metabólico (SM). La idea del mismo es proporcionar la evidencia para relacionar biológicamente los estresores psicosociales de la vida diaria y la enfermedad cardiaca.
El estrés en el trabajo se ha relacionado con la enfermedad coronaria en varios estudios, pero los mecanismos biológicos no están claros. El SM es un grupo de factores de riesgo que incrementan el riesgo de la enfermedad coronaria y la diabetes tipo 2 (DBT 2). Este síndrome está definido por tres o más de los siguientes factores: la obesidad abdominal, el aumento de los triglicéridos y el HDL bajo, la hipertensión arterial, la resistencia a la insulina medida por un aumento de la glucosa en ayunas.
El estudio reclutó 10308 hombres y mujeres de 35 a 55 años y se desarrolló en 5 fases a lo largo de 14 años. La fase 1 fue la de reclutamiento, las fases 2 y 4 eran en base a cuestionarios y la 3 y 5 incluían examen físico.
El SM fue definido por la presencia de tres o más factores de riesgo y el estrés laboral por medio de cuestionarios que se evaluaban en base a la demanda en el trabajo, si eran altas y si estaban relacionadas con la jerarquía del puesto y la toma de decisiones.
Al fin de la fase 5 los que completaron los datos clínicos sobre los indicadores del síndrome metabólico fueron el 75% (fallecieron en el transcurso del estudio 488 sujetos).
Los empleados con estrés crónico tuvieron más del doble de probabilidades de SM que los que no lo tenían (Odds Ratio 2,25, IC95% 1,31 a 3,85).
El estudio concluye que el estrés en el trabajo está asociado con la enfermedad coronaria aunque los mecanismos biológicos en esta asociación todavía no son claros.
Además, el síndrome metabólico tiene una relación inversa importante con la posición social y que hay una relación "dosis-respuesta" entre la exposición al estrés del trabajo y el síndrome metabólico.
Provee evidencia de una relación plausible entre los mecanismos del estrés psicosocial, más los estresores de la vida diaria, con la enfermedad coronaria y que parte de este gradiente social en el síndrome metabólico estaría explicado por la exposición crónica al estrés laboral

* ver glosario

Sergio A. Boero [Médico del Servicio de Clínica Medica, Hospital "Felipe Glasman" Bahia Blanca.]