# Cálculo del tamaño de muestra. Métodos secuenciales

*Preparado por Luis M. Molinero (Alce Ingeniería)*
CorreoE: **bioestadistica@alceingenieria.net**

**Artículo en formato PDF**

*Noviembre 2002*
www.seh−lelha.org/stat1.htm

En la planificación de cualquier trabajo de investigación surgirá indefectiblemente la pregunta "¿qué tamaño de muestra necesito para verificar la hipótesis planteada?", y si no ha sido así cuando vayamos a publicar el resultado de nuestro trabajo y procedamos a revisar la lista de ítems sugerida en la declaración CONSORT, nos encontraremos con el punto 7 que bajo el título "Tamaño de muestra" nos indica: *Especificar cómo fue determinado el tamaño de muestra y en su caso explicación de los análisis intermedios efectuados, así como reglas para detener el estudio o experimento*.

Aunque el razonamiento para la predeterminación del tamaño de muestra es tremendamente sencillo, y a pesar de que existen multitud de tablas publicadas y de programas para su cálculo, por algún extraño motivo muchos investigadores consideran la predeterminación del tamaño de muestra una tarea de "expertos" en estadística, lo que como veremos no tiene ningún sentido, pues la información más importante para ese cálculo se basa en conocer ciertos datos del proceso que se va a estudiar.

Como todo el mundo sabe, en un estudio comparativo podemos cometer dos tipos de errores, un error tipo I o $\alpha$, que ocurre cuando se afirma que existe diferencia y en realidad ésta es cero, y un error tipo II o $\beta$, que consiste en declarar que no hemos encontrado diferencias estadísticamente significativas cuando sí que son diferentes los dos grupos. Obviamente la realidad no la conocemos, y precisamente vamos a efectuar un trabajo para intentar saber más sobre ésta. Es habitual fijar de antemano la probabilidad de cometer un error de tipo I en una valor pequeño, normalmente inferior a 0.05. Uno de los problemas del contraste estadístico de hipótesis es que por pequeña que sea una diferencia ésta será estadísticamente significativa siempre que el tamaño de muestra sea suficientemente grande, de ahí el interés del concepto de relevancia clínica de una diferencia observada. Dado que al investigador lo que le interesa es encontrar diferencias con una magnitud de cierta importancia práctica y dado que el coste de un estudio aumenta con el tamaño de muestra, o lo que es lo mismo disminuye su viabilidad, ese orden de magnitud de la diferencia mínima que deseamos detectar permitirá acotar el tamaño de muestra necesario para nuestro estudio. La declaración de principios es que buscamos un tamaño de muestra lo más pequeño posible, pero no tanto que no seamos capaces de detectar una diferencia de una magnitud tal que ya empieza a tener interés práctico, es decir que si la observásemos en nuestro experimento deseamos tener un tamaño de muestra suficiente para poder afirmar que es estadísticamente significativa.

No vamos a entrar ahora a profundizar en la matemática del cálculo del tamaño de muestra, que se describe en prácticamente todos los libros de estadística, cuya fórmula de cálculo concreto dependerá del parámetro que se va a estimar (una proporción, una media, un coeficiente de correlación, etc) y del modelo probabilístico con el que se supone se distribuye ese parámetro, obteniéndose finalmente una fórmula que depende del error $\alpha$ prefijado, de la mínima diferencia *D* entre los parámetros que se considera de importancia práctica, de la probabilidad $\beta$ de no detectar esa diferencia, o su complementaria $1-\beta$ la probabilidad de detectarla o potencia de la prueba, y de la variabilidad de los datos expresada en función de la desviación típica *s*. Expresado de forma sucinta:

$$Tamaño = f(\alpha, \beta, D, s^2)$$

Donde *f( )* indica *función de ( )*

Por tradición $\alpha$ se suele fijar en 0.05 y la potencia de la prueba $1-\beta$ entre 0.8 y 0.9, y salvo que tengamos alguna razón poderosa para cambiarlo más nos vale no luchar con una tradición tan sólidamente asentada en el ámbito editorial científico.

En el caso de que estemos **comparando dos proporciones**, la varianza se puede expresar en función de esas proporciones, con lo que podemos formular el tamaño de muestra necesario de la siguiente manera

$$Tamaño = f(\alpha, \beta, P_1, P_2)$$

El problema radica en que para determinar el tamaño de muestra necesitamos conocer *P1* y *P2*, las proporciones en los dos grupos, que es precisamente lo que deseamos saber y para lo que pensamos efectuar un trabajo de investigación. Hablando coloquialmente, es la pescadilla que se muerde la cola, y sólo se puede resolver efectuando suposiciones, de ahí que quizás por no existir una solución "automática" es por lo que se considera, inmerecidamente, y en tantas ocasiones una tarea de "gurus". La forma habitual de proceder consiste en suponer un orden de magnitud de la tasa de respuesta en el grupo de control *P1*, basada en la experiencia previa, en la literatura, en un estudio piloto o simplemente en la intuición, y postular qué diferencia *D* en esa respuesta se puede empezar a considerar ya de interés, de tal manera que *P2=P1+D*. A partir de esos datos se calcula ya el tamaño de muestra necesario.

Es conveniente analizar en qué medida un valor diferente, pero también posible, de la tasa de respuesta conduce a otros tamaños de muestra, y sopesar así un rango de posibles tamaños junto con las restricciones logísticas y económicas, las cuales suelen tener la última palabra al respecto. Es habitual también prever que puede haber pérdidas de casos a lo largo del estudio, y de acuerdo a experiencias previas sobredimensionar inicialmente ese tamaño de muestra para garantizar que el tamaño al final del estudio no sea menor que el inicialmente previsto.

Para la **comparación de medias de datos cuantitativos**, además de fijar el orden de magnitud de la diferencia *D* que se considera ya de cierta importancia y que queremos garantizar que seremos capaces de detectar con una probabilidad entre 0.8 a 0.9 (potencia de la prueba), es preciso además conocer el valor de la desviación típica *s* que esperamos para nuestros datos, siendo éste el punto más complicado. También se puede acudir a fijar la relación entre esa diferencia y la desviación típica, el cociente *d=D/s*, cociente que se conoce con el nombre *d Cohen*, y que se suele utilizar para tabular posibles tamaños de muestra en función de ese valor *d*. Aquí el cuello de botella se encuentra, sobre todo en un estudio novedoso sin información previa, en cómo tener idea de una estimación sensata del valor de la desviación típica. Una alternativa posible es reestimar el tamaño de muestra una vez que hemos recogido ya parte de los datos, lo que nos permite obtener una estimación más real de la desviación típica. Este procedimiento nos lleva a técnicas de análisis secuencial que se comentan más adelante.

A la hora de determinar el tamaño de muestra necesario hay que tener en cuenta también el tipo de diseño y el tipo de muestreo utilizado, ya que éste condiciona la fórmula que se utilizará para calcular el error estándar de la estimación, y dado que manipulando esa fórmula se obtiene el valor de *N*, también éste será diferente según sea el diseño del estudio. Así, entre otros casos, habrá que tener en cuenta si se trata de una comparación de muestras independientes o pareadas, si el esquema del muestreo es aleatorio simple (el más utilizado en los ensayos clínicos) o si se empleó un método de muestreo diferente (por ejemplo un muestreo estratificado o un muestreo por conglomerados).

Como hemos visto el principal escollo que se presenta a la hora de predeterminar el tamaño de muestra necesario para un estudio, una vez que disponemos de las fórmulas, las tablas o el programa adecuado, lo

encontramos en que hay que aventurar valores de parámetros que lamentablemente van a ser a su vez calculados en el propio estudio y es éste un círculo vicioso que no tiene salida, a pesar de que algunos autores hayan propuesto insistentemente soluciones algo descabelladas. Así resulta curiosa la afirmación sostenida por algunos de que para estimar una proporción desconocida, con una precisión dada, el tamaño de muestra mínimo necesario se obtiene suponiendo un valor de *p=0.5*, basándose en que para estimar una proporción *P* con margen de tolerancia *D* la fórmula que proporciona el tamaño de muestra es:

$$n = \frac{Z^2_{1-\alpha/2} \cdot P \cdot (1-P)}{D^2}$$

donde Z=1.96 para $\alpha$=0.05.

Para *D* fijo esa fórmula toma su valor máximo con *P=0.5*. Pero *D* es la tolerancia en la estimación de la proporción y está claro que la magnitud de esa tolerancia no se puede fijar si no tenemos alguna idea respecto a la proporción a estimar. Un margen de tolerancia del 1% puede ser aceptable en la estimación de un porcentaje del 50%, o por ejemplo en un porcentaje del 20%, es decir que el intervalo de confianza de la estimación estaría en este último caso entre el 19% y el 21%. Pero esa misma tolerancia es probablemente inadmisible para estimar un porcentaje del 2%, ya que entonces el margen absoluto del 1% constituye la mitad del valor estimado. El propio sentido común nos dice que para estimar sucesos infrecuentes necesitaremos tamaños de muestra mayores que para estimar sucesos frecuentes.

## Métodos secuenciales

Hasta aquí se ha considerado la metodología clásica del contraste de hipótesis, en la que se fija el tamaño de muestra al principio del estudio, ya sea mediante un cálculo basado en los razonamientos expuestos anteriormente o por otros motivos más prosaicos, aunque luego para su publicación se "ajusten" las cifras a alguna de las fórmulas anteriores. Pero existe una alternativa, que se conoce con el nombre de análisis secuencial en la que se efectúa los cálculos a medida que se van recogiendo casos y se compara el resultado obtenido hasta ese momento con dos umbrales, superados los cuales se detiene el experimento. Supongamos que se está comparando dos tratamientos y que el resultado analizado es dicotómico, por ejemplo para simplificar supongamos curación sí o no. En un ensayo secuencial fijamos entonces un umbral superior para la diferencia de proporciones entre la tasa en el grupo de nuevo tratamiento y la tasa en el grupo control, superado el cual se detiene el estudio considerando que hay ya suficiente evidencia para aceptar el nuevo tratamiento como mejor. Asimismo habrá un umbral inferior, por debajo del cual se considerará probado que el nuevo tratamiento es peor y se detendrá también el experimento. Esta metodología permite obtener conclusiones durante el desarrollo del estudio, sin necesidad de esperar al final del mismo, por lo que habitualmente se puede llegar a una conclusión más rápidamente que en un diseño clásico, lo que en determinadas situaciones lo hace muy deseable ya que puede evitar a tiempo que los pacientes estén recibiendo un tratamiento mucho menos eficaz. No es de extrañar que los métodos secuenciales se utilicen asiduamente en la industria en ensayos destructivos, en los que la pieza que se analiza queda inoperativa después del proceso de análisis.

En la metodología estadística secuencial para mantener la probabilidad global de error tipo I en un valor inferior al prefijado (0.05), es necesario utilizar un nivel de significación menor para cada paso. Se trata de una problemática similar a la comentada para el contraste múltiple de hipótesis. La regla de parada puede ser detener el estudio si los datos acumulados hasta ese momento muestran una diferencia estadísticamente significativa a un nivel de probabilidad tal que se garantiza el objetivo global de $\alpha$, y en caso contrario continuar con el estudio hasta llegar a un tamaño máximo *N* momento en el que también se detendrá el estudio y se aceptará la hipótesis de igualdad. El nivel de probabilidad a utilizar en cada paso depende del valor de *N*, cuanto mayor sea éste menor debe ser ese valor crítico.

Actualmente los métodos secuenciales ha cobrado auge en epidemiología genética, fundamentalmente en la metodología que se conoce como "sequential linkage analysis". También en muchos ensayos clínicos se lleva a cabo de forma habitual análisis intermedios de los datos, con el fin de determinar si los resultados obtenidos hasta ese momento hacen inaceptable, desde un punto de vista ético, la continuidad del estudio, porque se ha encontrado una diferencia de efectos inesperadamente grande o porque por ejemplo la tasa de efectos secundarios es anormalmente alta. En estos casos es necesario también prefijar las condiciones de parada del estudio.

Otra cuestión interesante que podemos plantearnos es qué influencia tiene el tamaño de un estudio en la opinión clínica general. El intervalo de confianza de la magnitud de un efecto depende del tamaño de muestra y de la varianza, pero lo que en general quizás llama más la atención es precisamente el tamaño de muestra. ¿Hasta qué punto sería recibido con escepticismo un ensayo en el que como consecuencia de la detección de un gran diferencia en el efecto éste se detiene a la mitad de su desarrollo, cuando llevamos incluidos la mitad de los pacientes previstos?. Si existiera una tendencia a sobrevalorar un gran tamaño de muestra como garante de evidencia de un resultado, sería necesario incluir en el cálculo de las reglas de parada no sólo criterios puramente estadísticos sino además otros que cuantificasen de alguna manera el "peso" de la evidencia necesaria para alterar la práctica clínica, lo que precisaría de un enfoque bayesiano de la cuestión. Sin embargo, ahora que están tan de moda los macro−estudios, hemos de tener bien claro que los grandes estudios sólo son necesarios para demostrar pequeños efectos o diferencias pequeñas en los efectos y que aunque las grandes cifras del efectivo de muestra de esos trabajos nos impresionen, el tamaño no lo es todo, ya que en la calidad de los resultados también influye la calidad y características del diseño y ésta siempre es mucho más difícil de garantizar en los grandes estudios en los que intervienen centros muy diversos que en estudios más pequeños y mucho mejor controlados.

## Enlaces de interés

- PS: Power and Sample Size software
  Dupont WD and Plummer WD: PS power and sample size program available for free on the Internet. Controlled Clin Trials,1997;18:274
  PS is an interactive program for performing power and sample size calculations. The program runs on the Windows 95, Windows 98, Windows NT, and Windows 2000 operating systems. It can be used for studies with dichotomous, continuous, or survival response measures. The alternative hypothesis of interest may be specified either in terms of differing response rates, means, or survival times, or in terms of relative risks or odds ratios. Studies with dichotomous or continuous outcomes may involve either a matched or independent study design.

- 1) Sample Size Calculation
  2) Wald's Sequential Probability Ratios
  SISA Simple Interactive Statistical Analysis
  Daan Uitenbroek
  Hilversum
  The Netherlands

- Power, Sample Size and Experimental Design Calculators...
  Enlaces recopilados por John C. Pezzullo
  Professor in the Departments of Pharmacology and Biostatistics
  Georgetown University, Washington DC