Revista Mexicana de Ciencias Agrícolas Vol.4 Núm. 1 1 de enero - 14 de febrero, 2013 p. 33-47

Tamaño de muestra para estimar expresión genética de plantas transgénicas usando pruebas de grupo*

Sample size for estimating gene expression of transgenic plants using group tests

Osval Antonio Montesinos-López[1§], Abelardo Montesinos-López[2], Ignacio Luna-Espinoza[3] y Jesús Erasmo Montesinos-López[2]

[1]Facultad de Telemática, Universidad de Colima. Bernal Díaz del Castillo Núm. 340, Villas San Sebastián, 28045. Colima, México. [2]Departamento de Estadística, Centro de Investigación en Matemáticas (CIMAT), Guanajuato, México. (aml_uach2004@yahoo.com.mx; shumyto@hotmail.com). [3]Universidad del Istmo-Campus Ixtepec. Cd. Ixtepec, 70110, Oaxaca, México. (iluna@bianni.unistmo.edu.mx). [§]Autor para correspondencia: oamontes2@hotmail.com.

Resumen

En las regiones sur-este y centro-oeste de México se ha detectado la presencia de maíz transgénico (GM), aun cuando los efectos de la polinización cruzada entre este maíz y variedades criollas o silvestres, como *Tripsacum* y teocintle, son desconocidos. Por esta razón es necesario detectar la presencia de plantas transgénicas y estimar la expresión genética media de los transgenes en los cultivos normales. Sin embargo, hacer un análisis de cada una de las plantas consumiría mucho tiempo y dinero. Una alternativa para reducir costos es utilizar la prueba de grupos. Esta técnica analiza grupos que contienen tejidos de varias plantas sin inspeccionarlas individualmente, manteniendo niveles aceptables de exactitud pero a bajo costo. Cuando la prueba de grupos se utiliza para estimar expresión genética media, es importante determinar el tamaño de muestra, número de grupos, para realizar el proceso de estimación. En este contexto y bajo condiciones de normalidad, este trabajo presenta dos procedimientos, computacional y analítico, para estimar la expresión genética media de maíz GM y se proporcionan ejemplos para mostrar la aplicación de los métodos propuestos. Adicionalmente, mediante simulación se reprodujeron diversas circunstancias que un investigador puede encontrar y se proporciona un algoritmo computacional en el programa estadístico R (R

Abstract

In the south-east and west-center of Mexico, the presence of transgenic maize has been detected (GM), even though the effects of cross-pollination between this maize and landraces such as *Tripsacum* and teosinte are unknown so far. It is therefore necessary to detect the presence of transgenic plants and to estimate the average gene expression of transgenes in regular crops. However, an analysis of each and every one of the plants would consume a lot of time and money as well. An alternative to reduce costs is to use test groups. This technique analyzes groups that possess tissues of several plants without individually inspecting them, keeping acceptable levels of accuracy at a lower cost. When a test group is used to estimate the average genetic expression it is important to determine the sample size and the number of groups for the estimation process. In this context and under conditions of normality, this paper presents two procedures, computational and analytical, for estimating the average gene expression of GM maize, providing examples to show the application of the proposed methods. Additionally, through simulation we reproduced several circumstances that a researcher might find, providing a computational algorithm in the statistical program R (R Development Core Team 2007) to create other scenarios. Both procedures ensure that, W confidence interval amplitude is lower or

---

* Recibido: julio de 2012
Aceptado: enero de 2013

Development Core Team 2007) para crear otros escenarios. Ambos procedimientos garantizan que la amplitud del intervalo de confianza W sea menor o igual que la amplitud deseada ω con probabilidad γ. Esto se logra porque los dos métodos consideran la aleatoriedad de la varianza muestral. Sin embargo, se recomienda el uso de la fórmula propuesta, procedimiento analítico, porque es preciso y sencillo de usar.

Palabras clave: intervalo de confianza, expresión genética media, normalidad.

Introducción

La diseminación de transgenes en los cultivos normales o parientes silvestres es un riesgo inherente en la agricultura. En México, un país que alberga más 60% de la variedad genética del maíz (*Zea mays* L.) (Piñeyro-Nelson *et al*., 2009), los efectos por diseminar genes de plantas de maíz genéticamente modificadas (GM), son desconocidos, aún cuando investigaciones recientes testifican la presencia de transgenes de maíz en las regiones sur-este y centro-oeste del país (Dyer *et al*., 2009; Piñeyro-Nelson *et al*., 2009). Ante este escenario de preocupación es importante detectar la presencia y expresión de trangenes en los cultivos de maíz.

La detección de plantas transgénicas involucra la evaluación de compuestos orgánicos o el análisis del metabolismo de cada uno de los elementos de la población. Sin embargo, efectuar una inspección o análisis a cada elemento consumiría grandes cantidades de recursos económicos y materiales. Una alternativa para disminuir costos es efectuar muestreo en grupos antes de realizar las mediciones analíticas, esto reduciría el total de análisis. Además, usar muestreo en grupos para caracterizar poblaciones no solo es más eficiente en términos económicos, sino también pueden realizarse estimaciones más precisas y menos sesgadas que las obtenidas con muestras individuales (Caudill, 2010).

El muestreo en grupos o prueba de grupos, propuesto por Dorfman (1943), consiste en formar g grupos de tamaño k con muestras individuales. Esta forma de agrupar elementos puede usarse para: a) reducir la variación biológica; b) disminuir costos vía la reducción del total de pruebas de laboratorio; y c) que la disponibilidad de muestras limitadas no sea un problema. Debido al ahorro significativo de tiempo y dinero por utilizar esta técnica, su empleo se incrementa día a día, usándose para detectar enfermedades al donar sangre

equal to the desired amplitude ω with γ probability. This is achieved because both methods consider the randomness of the sampling variance. However, we recommend the use of the formula proposed and analytical procedure, because it is accurate and easy to use.

Key words: confidence interval, average gene expression, normality.

Introduction

The spread of transgenes in regular crops or wild relatives is an inherent risk in agriculture. In Mexico a country that is home to over 60% of the genetic diversity of maize (*Zea mays* L.) (Piñeyro-Nelson *et al*., 2009), the effects of spreading plant genes of genetically modified maize (GM) are unknown, even though recent research testify the presence of transgenes in maize in the southern-east and center-west (Dyer *et al*., 2009; Piñeyro-Nelson *et al*., 2009). Given this scenario of concern it is important to detect the presence and expression of transgenes in maize crops.

Detection of transgenic plants involves evaluation of organic compounds or metabolic analysis of each of the elements of the population. However, inspecting or testing each item would consume large amounts of financial resources and materials. An alternative to reduce costs is to perform group samplings before the analytical measurements; thus, reducing the total of analysis. Also, using group sampling to characterize populations is not only more efficient in economic terms, but also the estimates can be made even more accurate and less biased than those obtained with individual samples (Caudill, 2010).

Group sampling or cluster sampling proposed by Dorfman (1943) consists of forming g groups of size k with individual samples. This form of grouping items can be used to: a) reduce biological variation, b) reduce costs through reduction of total laboratory testing, and c) the limited sample availability is not a problem. Due to significant savings of time and money by using this technique, its use is increasing day by day, being used to detect disease by donating blood (Dodd *et al*., 2002), drugs (Remlinger *et al*., 2006), transgenic plants (Hernandez-Suarez *et al*., 2008; Montesinos-López *et al*., 2010), to estimate the prevalence of human diseases (Verstraeten *et al*., 1998), plant diseases (Tebbs and Bilder, 2004) and animals (Peck, 2006).

(Dodd *et al.*, 2002), drogas (Remlinger *et al.*, 2006), plantas transgénicas (Hernández-Suárez *et al.*, 2008; Montesinos-López *et al.*, 2010), estimar la prevalencia de enfermedades humanas (Verstraeten *et al.*, 1998), enfermedades de plantas (Tebbs y Bilder, 2004) y animales (Peck, 2006).

En la prueba de grupos, grupos de individuos son caracterizados en lugar de elementos individuales. De acuerdo con el supuesto del promedio biológico, la medición en una muestra en grupo es comparable con la media aritmética de los niveles individuales de los elementos que conforman el mismo grupo (Mary-Huard *et al.*, 2007; Caudill, 2010). Por lo tanto, si una cantidad medible se distribuye normalmente, la distribución de las medidas del muestreo en grupos también tendrán distribución normal con la misma media pero con varianza reducida proporcionalmente por el número de elementos en el grupo.

Cuando el objetivo es estimar la expresión media de un gen que produce plantas GM en la población, es importante diseñar un experimento que garantice el tamaño de muestra apropiado para asegurar intervalos de confianza cortos (Shaarschimidt, 2007). Un tamaño de muestra pequeño no garantiza buena precisión en la estimación del parámetro de interés, mientras un tamaño de muestra grande es un derroche innecesario de recursos (Wang *et al.*, 2005). En el contexto de los experimentos de microarreglos y considerando variables aleatorias continuas, Kendziorski *et al.* (2003), Dobbin y Simon (2005) y Zhang y Gant (2005) determinaron los tamaños de muestra bajo el muestreo en grupos. Kendziorski *et al.* (2003) proporcionó una fórmula que determina el número de grupos para estimar la expresión de genes y establecer intervalos de confianza (ICs), comparando los resultados con aquellos obtenidos sin utilizar prueba en grupos. Sin embargo, el trabajo ignora la naturaleza estocástica de la amplitud del intervalo de confianza (IC).

Por otro lado, Dobbin y Simon (2005) y Zhang y Gant (2005) obtuvieron el tamaño de muestra bajo el enfoque de potencia en muestreo en grupos, razón por la cual los resultados no son apropiados para producir estimaciones precisas de la media u otro parámetro de interés.

Típicamente se han propuesto fórmulas para determinar tamaños de muestra bajo condiciones de potencia. Este enfoque es adecuado cuando se hacen pruebas de hipótesis, reportando los resultados en términos de p-values. Sin embargo, para hacer inferencias actualmente se ha incrementado el uso de ICs en lugar de pruebas de hipótesis

In the test groups, groups of individuals are characterized instead of individual elements. Under the assumption of average biological measurements in a sample group is comparable to the average of the individual levels of the elements of the same group (Mary-Huard *et al.*, 2007; Caudill, 2010). Therefore, if a measurable quantity is normally distributed, the distribution of the measurements of the sampling groups will also be normally distributed with the same mean variance but, proportionately reduced by the number of elements in the group itself.

When the objective is to estimate the mean expression of a gene that produces GM plants in the population, it is important to design an experiment to ensure proper sample size to ensure shorter confidence intervals (Shaarschimidt, 2007). A small sample size does not guarantee good accuracy in estimating the parameter of interest, while a large sample size is an unnecessary waste of resources (Wang *et al.*, 2005). In the context of microarray experiments and considering continuous random variables, Kendziorski *et al.* (2003), Dobbin and Simon (2005) and Zhang and Gant (2005) determined the sample sizes under cluster sampling. Kendziorski *et al.* (2003) provided a formula that determines the number of clusters to estimate the gene expression and establish confidence intervals (CIs), comparing the results with those obtained without using test groups. However, the work ignores the stochastic nature of the amplitude of the confidence interval (CI).

On the other hand, Dobbin and Simon (2005) and Zhang and Gant (2005) obtained the sample size under the power approach in the groups sampling, for this reason, the results are not suitable to produce accurate estimates of the mean or other parameter of interest.

Formulas have been proposed to determine sample sizes under power approach. This approach is appropriate when testing hypotheses, reporting results in terms of p-values. However, in order to make inferences actually increased the use of CIs rather than hypothesis testing (Pan and Kupper, 1999). In agricultural studies estimating parameters via IC is important because often the main goal is to estimate the magnitude of the effect of interest, and not only decide whether the treatment effects are statistically different or not.

A hypothesis test points if an effect is significant without providing a precise characterization of the effect being tested in the null hypothesis. Using CIs not only ensures that the magnitude of the effect can be assessed, but also to

(Pan y Kupper, 1999). En los estudios agrícolas la estimación de parámetros vía IC es importante porque frecuentemente el objetivo principal es estimar la magnitud del efecto de interés, y no sólo decidir si los efectos de los tratamientos son estadísticamente diferentes.

Una prueba de hipótesis puntualiza si un efecto es significativo sin proporcionar una caracterización precisa del efecto que está siendo probado en la hipótesis nula. El uso de ICs asegura no solo que la magnitud del efecto pueda evaluarse, sino también que el efecto en estudio pueda ser identificado fácilmente por el lector. Además, los ICs transmiten información para que la magnitud del efecto pueda determinarse a partir de los datos disponibles (Beal, 1989). Por tales razones se ha puesto atención al calcular tamaños de muestras apropiados para realizar inferencias basadas en ICs. Este enfoque de estimación se ha denominado precisión en la estimación de parámetros (PEP) porque cuando la amplitud del IC con una confiabilidad de $(1 - \alpha)$ 100% decrece, la precisión esperada de la estimación aumenta (Kelley y Maxwell, 2003; Kelley y Rausch, 2006; Kelley, 2007).

Cuando se determina el tamaño de muestra se requieren los valores de algunos parámetros. En la práctica éstos son desconocidos y usualmente se estiman de la literatura o estudios previos. Éstas estimaciones son consideradas como los verdaderos valores de los parámetros, trayendo como consecuencia que con el tamaño de muestra calculado no se logre la precisión deseada en el IC (Wang *et al.*, 2005). Para tener en cuenta la incertidumbre inducida por el error de muestreo, Kelley (2007) y Kupper y Hafner (1989) señalaron que la naturaleza estocástica de la amplitud del IC debería considerarse para no subestimar el tamaño de muestra requerido. Así, bajo el modelo de Dorfman, Montesinos-López *et al.* (2010) propuso un procedimiento que determina el tamaño de muestra para estimar la proporción de plantas transgénicas, asegurando que la amplitud W del IC sea más estrecha que el valor deseado ω. Sin embargo, este método no proporciona una solución analítica.

Por tales motivos, bajo el contexto de la prueba de grupos con prueba perfecta y tamaño de grupo fijo, el objetivo de esta investigación fue proponer un método analítico que determine el tamaño de muestra dado en términos del número de grupos requeridos para estimar por intervalo la expresión media de un gen, asegurando ICs estrechos. La precisión en la estimación de la media se logra porque se considera la aleatoriedad de la amplitud del IC. Además se presenta

study the effect that can be easily identified by the reader. In addition, CIs transmit information to the effect size that can be determined from the available data (Beal, 1989). For these reasons, attention has been paid to calculate appropriate sample sizes to make inferences based on CIs. This approach has been termed accuracy in parameter estimation (PEP) because when the CI with a reliability of $(1 - \alpha)$ 100% decreases, the expected accuracy of the estimation increases (Kelley and Maxwell, 2003; Kelley and Rausch, 2006; Kelly, 2007).

When determining the sample size, values of some parameters are required. In practice these are usually unknown and are estimated from the literature or previous studies. These estimates are considered to be the true values of the parameters, consequently resulting in the calculated sample size not achieving the desired accuracy in the CI (Wang *et al.*, 2005). To account for the uncertainty induced by the sampling error, Kelley (2007) and Kupper and Hafner (1989) noted that, the stochastic nature of the CI should be considered not to underestimate the required sample size. Thus, under the model of Dorfman, Montesinos-López *et al.* (2010) proposed a method which determines the sample size to estimate the proportion of transgenic plants, ensuring that the W amplitude of CI is narrower than the desired value. However, this method does not provide an analytical solution.

For these reasons, in the context of proving perfect test groups with fixed group size, the objective of this research was to propose an analytical method to determine the sample size given in terms of the number of groups per interval required for estimating the mean expression of a gene, ensuring close CIs. The estimation accuracy is achieved because the average is considered the randomness of the CI. It also presents a computational algorithm in the statistical program free to use and free distribution R (R Development Core Team 2007) to get the results, so that researchers can play other scenarios.

Materials and methods

Being X the amount getting determined in the population or experiment; *i.e.*, the level of expression of a gene. Let x_i denote the value of the X variable in the i element of the population of interest. It is assumed that all the $x_i s$ in the population are independent, normally distributed with μ means, σ^2 variance, denoted by $x_i \sim N(\mu, \sigma^2)$, for every i.

un algoritmo computacional en el programa estadístico de uso libre y distribución gratuita R (R Development Core Team 2007) para obtener los resultados, de tal forma que los investigadores puedan reproducir otros escenarios.

Materiales y métodos

Sea X la cantidad medida que está siendo determinada en la población o experimento; es decir, el nivel de expresión de un gen. Permita que x_i denote el valor de la variable X en el elemento i de la población de interés. Se asume que todas las $x_i s$ en la población son independientes, normalmente distribuidas con media μ y varianza σ^2, denotado por $x_i \sim N(\mu, \sigma^2)$, para toda i.

Los elementos de la población son seleccionados aleatoriamente y a cada uno se le extrae una muestra de tejido. Un grupo de muestras de tejidos es formado congregando k muestras de tejidos de elementos individuales, los cuales son seleccionados aleatoriamente (Zhang y Gant, 2005). Así se forman g grupos de k elementos cada uno, donde g, k son enteros positivos y $n= g, k. n$ es el número total de muestras individuales (o sujetos), aunque en este caso no se realizan las mediciones de las muestras individuales. En su lugar, estas muestras individuales son agrupadas en g grupos con k muestras en cada grupo y m mediciones (réplicas) se efectúan en cada grupo de muestras de tejidos. Por lo tanto, m es el número de réplicas técnicas de la medida en cada grupo (Zhang y Gant, 2005; Caudill, 2010). Note que si k =1, el experimento es equivalente a no realizar grupos de muestras de tejidos; y si m=1 , no existen réplicas.

Bajo el supuesto básico del promedio biológico, el resultado por agrupar k muestras de tejidos en proporciones iguales es que el valor de k en cada grupo es el promedio de los elementos que conforman este mismo grupo (Zhang y Gant, 2005), Además, $\bar{x}= k\text{-}1\ \Sigma_{i=1}^{k} x_i$ tiene distribución normal con media μ y varianza $k^{-1}\sigma^2$ para cada grupo de la población (Zhang y Gant, 2005). En este artículo sólo se discuten muestras de grupos con contribuciones individuales iguales. Aunque se pueden formar grupos con contribuciones desiguales de las muestras individuales, tal diseño es generalmente menos efectivo que con contribuciones iguales (Peng *et al.,* 2003).

De acuerdo con Zhang y Gant (2005), cuando se toma una medida sobre un grupo p, el valor medido es $y_{p,r}=\bar{x}_p+\varepsilon_r$, donde p indica el grupo, r hace referencia a las mediciones y ε_r son

The elements of the population are randomly selected and each will take a sample of tissue. A group of tissue samples is formed gathering k tissue samples from individual elements, which are randomly selected (Zhang and Gant, 2005). Thus forming g groups of k elements each, where g, k are positive integers and n= gk.n is the total number of individual samples (or subject); although in this case the measurements are not made of individual samples. Instead, these individual samples are grouped in g groups with k samples for each group and m measurements (replicas) are effected in each group of tissue samples. Therefore, m is the number of technique replicas of the measurement in each group (Zhang and Gant, 2005; Caudill, 2010). Note that if k=1, the experiment is not equivalent to perform groups of tissue samples; and if m=1 there are no replicas.

Under the basic assumption of biological averaging the result from grouping k tissue samples in equal proportions is that the value of k in each group is the average of the elements of this group (Zhang and Gant, 2005), also $\bar{x}= k\text{-}1\ \Sigma_{i=1}^{k} x_i$ has a normal distribution with mean μ and $k^{-1}\sigma^2$ variance for each population group (Zhang and Gant, 2005). This paper discusses only group samples with equal individual contributions. Even though groups with unequal contributions can be created of individual samples, such design is generally less effective than equal contributions (Peng *et al*., 2003).

According to Gant and Zhang (2005), when a measurement is taken on a p group, the measured value is $y_{p,r}=\bar{x}_p+\varepsilon_r$, where p indicates the group, r refers to the measurements and ε_r are random errors, which are assumed to be independent of normal distribution $\varepsilon_r \sim N(0,\sigma_\varepsilon^2)$. Henceforth σ_ε^2 will be called the technical variance and σ^2 the biological variance of the population. Measurements of the g groups on the experiment results. Thus we have $y_{p,r}$, for p=1,…, r=1,…, m: and g the number of groups formed from the sample of the population (Zhang and Gant, 2005). The objective is to make inferences about the properties of the population based on the available data.

It can be shown that $\bar{Y}=(mg)^{-1}\Sigma_{p=1}^{g}\Sigma_{r=1}^{m} y_{p,r}$ is a direct estimator of μ (Zhang y Gant, 2005; Caudill, 2010), with variance $\sigma_{\bar{Y}}^2= g^{-1}(k^{-1}\sigma^2 + m^{-1}\sigma_\in^2)= g^{-1}\sigma_p^2$, where $\sigma_p^2= k^{-1}\sigma^2 + m^{-1}\sigma_\in^2$, and $s_{\bar{Y}}^2= g^{-1}((g-1)^{-1}\ \Sigma_{p=1}^{g}(m^{-1}\Sigma_{r=1}^{m} y_{p,r} - \bar{Y})^2)= g^{-1}\sigma_p^2$ is a direct estimator of $\sigma_{\bar{Y}}^2$ (Zhang y Gant, 2005; Kendziorski *et al*., 2003), where $s_p^2=(g-1)^{-1}\ \Sigma_{p=1}^{g}(m^{-1}\Sigma_{r=1}^{m} y_{p,r} - \bar{Y})^2$. Therefore, the corresponding CI of Wald is errores aleatorios, los cuales se asume son independientes con distribución normal $\varepsilon_r \sim N(0,\sigma^2_\varepsilon)$. De aquí en adelante a σ^2_ε se le llamará la varianza técnica y σ^2 la varianza biológica de la población. Las mediciones sobre los g grupos son los resultados del experimento. Así se tiene $y_{p,r}$, para $p= 1,\dots, g$, $r= 1,\dots, m$; y g es el número de grupos formados a partir de una muestra de la población (Zhang y Gant, 2005). El propósito es realizar inferencias sobre las propiedades de la población con base en los datos disponibles.

Puede mostrarse que $\bar{Y}= (mg)^{-1}\Sigma^g_{p=1}\Sigma^m_{r=1}\, y_{p,r}$ es un estimador insesgado de μ (Zhang y Gant, 2005; Caudill, 2010), con varianza $\sigma^2_{\bar{Y}}= g^{-1}(k^{-1}\sigma^2 + m^{-1}\sigma^2_\epsilon)= g^{-1}\sigma^2_p$, donde $\sigma^2_p= k^{-1}\sigma^2 + m^{-1}\sigma^2_\epsilon$, y $s^2_{\bar{Y}}= g^{-1}((g - 1)^{-1}\, \Sigma^g_{p=1}(m^{-1}\Sigma^m_{r=1}\, y_{p,r} - \bar{Y})^2)= g^{-1}\sigma^2_p$ es un estimador insesgado de $\sigma^2_{\bar{Y}}$ (Zhang y Gant, 2005; Kendziorski *et al*., 2003), donde $s^2_p= (g - 1)^{-1}\, \Sigma^g_{p=1}(m^{-1}\Sigma^m_{r=1}\, y_{p,r} - \bar{Y})^2$. Por lo tanto, el correspondiente IC de Wald es:

$$[\mu_L= \bar{y} - Z_{1\text{-}\alpha/2}\, \sqrt{s^2_p / g},\ \ \mu_U= \bar{y} - Z_{1\text{-}\alpha/2}\sqrt{s^2_p / g}] \qquad (1)$$

Donde: $Z_{1\text{-}\alpha/2}$ es el cuantil $1\text{-}\alpha/2$ de la distribución normal estándar. La ecuación (1) es igual a la propuesta por Kendziorski *et al*. (2003), pero con la diferencia de que se usa $Z_{1\text{-}\alpha/2}$ en lugar del cuantil $1\text{-}\alpha/2$ de la distribución t-student con *g-1* grados de libertad. Esta sustitución no provoca problemas graves de subestimación. Por otro lado, este IC es fácil de calcular y permite derivar una fórmula cerrada para el tamaño de muestra, aunque cuando g es pequeño, la amplitud del IC es más grande.

Derivación del tamaño de muestra para estimar la media

La cantidad $Z_{1\text{-}\alpha/2}\sqrt{s^2_p/ g}$ en la ecuación (1), la cual es sumada y sustraída a la media observada $\bar{y}$, se define como $W/2$ (W es la amplitud total del IC). Los límites superior e inferior del IC están determinados por $W/2$. El grado de precisión del IC es el valor de más interés en trabajos con enfoques PEP. El valor W se fija a priori de acuerdo a la precisión deseada al estimar los parámetros.

La amplitud total del IC [Ecuación (1)] puede expresarse como $W= 2Z_{1\text{-}\alpha/2}\sqrt{s^2_p / g}$. Para estimar el número de grupos necesarios (tamaño de muestra) con el fin de estimar la media μ, dada una amplitud esperada ω (error), debe resolverse la Ecuación $W= 2Z_{1\text{-}\alpha/2}\sqrt{s^2_p / g}$ en términos de g(haciendo $W= \omega$), y la varianza muestral s^2_p debe reemplazarse por el valor poblacional σ^2_p, produciéndose así la siguiente fórmula:

$$g= w^{-2}2^2Z^2_{1\text{-}\alpha/2}\,\sigma^2_p= (w^{-1}2Z_{1\text{-}\alpha/2})^2\, (k^{-1}\sigma^2 + m^{-1}\sigma^2_\varepsilon) \qquad (2)$$

$$[\mu_L= \bar{y} - Z_{1\text{-}\alpha/2}\, \sqrt{s^2_p / g},\ \ \mu_U= \bar{y} - Z_{1\text{-}\alpha/2}\sqrt{s^2_p / g}] \qquad (1)$$

Where: $Z_{1\text{-}\alpha/2}$ is the quantile $1\text{-}\alpha/2$ of the standard normal distribution. Equation (1) is equal to that given by Kendziorski *et al*. (2003), but with the difference that it uses $Z_{1\text{-}\alpha/2}$ instead of the quantile $1\text{-}\alpha/2$ of the t-student distribution with g-1 degrees of liberty. This substitution does not cause serious understatement problems. Furthermore, this CI is easy to calculate and can derive a closed formula for the sample size, although when g is small, the CI is larger.

Sample size derivation for estimating the mean

The amount $Z_{1\text{-}\alpha/2}\sqrt{s^2_p / g}$ in equation (1), which is summed and subtracted to the y observed mean, is defined as W/2 (W is the total amplitude of CI). The upper and lower limits of CI are determined by W/2. The degree of accuracy of CI is the value of more interest in papers with PEP approaches. The W value is set *a priori* according to the desired precision in estimating the parameters.

The total amplitude of CI [Equation (1)] can be expressed as $W= 2Z_{1\text{-}\alpha/2}\sqrt{s^2_p/ g}$. To estimate the required number of groups (sample size) so as to estimate the mean μ, given an expected ω amplitude (Error), the equation $W= 2Z_{1\text{-}\alpha/2}\sqrt{s^2_p / g}$ must be solved in terms of g $W= \omega$) , and the sample variance s^2_p must be replaced with a population value σ^2_p, thereby producing the following formula:

$$g= w^{-2}2^2Z^2_{1\text{-}\alpha/2}\,\sigma^2_p= (w^{-1}2Z_{1\text{-}\alpha/2})^2\, (k^{-1}\sigma^2 + m^{-1}\sigma^2_\varepsilon) \qquad (2)$$

This formula can be used to estimate the required number of groups that were used in the estimation of the average expression of a gene, considering a fixed size group k, m measurements per group, and assuming that σ^2 and σ^2_ε are known. Note that if k=m=1, Equation (2) is reduced to the standard formula for estimating the mean under simple random sampling [$n=w^{-2}4Z^2_{1\text{-}\alpha/2}\sigma^2$]. However, in the equation (2) values σ^2 and σ^2_ε and are unknown, so that their respective estimators are used.

With the equation (2) the sample size is determined that yields a CI of W amplitude for estimating the mean, under the context of the test groups. However, there is no guarantee that the observed W amplitude is accurate for a particular CI, because σ^2 and σ^2_ε estimates are used. This implies that about 50% of the sampling distribution of W is lower than ω(Kelley and Maxwell, 2003; Kelley *et al*., 2003; Montesinos-López *et al*., 2010). Then we propose a formula for calculating sample sizes to ensure optimal CIs accurate enough.

Ésta fórmula puede usarse para estimar el número de grupos requeridos que se usaran en la estimación de la expresión media de un gen, considerando un tamaño de grupo fijo k, m mediciones por grupo, y asumiendo que σ^2 y σ^2_ε son conocidas. Note que si $k= m= 1$, la Ecuación (2) se reduce a la fórmula estándar para estimar la media bajo muestreo aleatorio simple [$n= w^{-2}4Z^2_{1-\alpha/2}\sigma^2$]. Sin embargo, en la ecuación (2) los valores de σ^2 y σ^2_ε son desconocidos, por lo que se usan sus respectivos estimadores.

Con la ecuación (2) se determina el tamaño de muestra que arroje un IC de amplitud W para estimar la media, bajo el contexto de la prueba en grupos. Sin embargo, no existe garantía de que la amplitud observada W sea precisa para algún IC en particular porque se usan estimaciones de σ^2 y σ^2_ε. Esto implica que aproximadamente 50% de la distribución muestral de W sea menor que ω (Kelley y Maxwell, 2003; Kelley *et al*., 2003; Montesinos-López *et al*., 2010). En seguida se propone una fórmula para calcular tamaños de muestra óptimos que garantizan ICs suficientemente precisos.

Procedimiento para calcular tamaños de muestra óptimos

La amplitud del IC para la media es $2Z_{1-\alpha/2}\sqrt{s^2_p / g}$, donde $s^2_p= (g -1)^{-1} \Sigma^g_{p=1} (m^{-1}\Sigma^m_{r=1}y_{p,r} - \bar{Y})^2$ W debería ser menor que un valor especifico ω con probabilidad γ. Así, siguiendo la lógica de Hahn y Meeker (1991) y Montesinos-López *et al*. (2011), para determinar un tamaño de muestra que asegure ICs estrechos, el tamaño de muestra óptimo es el valor entero más pequeño g_m tal que:

$$P\{W \leq \omega\} \geq \gamma$$

$$P\{2Z_{1-\alpha/2}\sqrt{s^2_p / g} \leq \omega\} \geq \gamma$$

$$P((gm - 1)s^2_p/ \sigma^2_p \leq gm(gm - 1)\omega^2 / 4Z^2_{1-\alpha/2}\sigma^2_p)) \geq \gamma, \sigma^2_p = (k^{-1}\sigma^2 + m^{-1}\sigma^2)$$

$$\text{así,...}P(X^2_{gm-1} \leq gm\,(gm - 1)\omega^2 / 4Z^2_{1-\alpha/2}\sigma^2_p)) = \gamma$$

$$\Leftrightarrow gm\,(gm - 1)\omega^2 / 4Z^2_{1-\alpha/2}\sigma^2_p)) = X^2_{gm-1,\gamma}$$

Por lo tanto, el número de grupos requeridos es

$$gm= 4Z^2_{1-\alpha/2}(k^{-1}\sigma^2 + m^{-1}\sigma^2)\,(X^2_{gm-1,\gamma}) / (\omega^2(gm - 1) \qquad (3)$$

Donde: γ es el grado de certeza (probabilidad requerida) para lograr que la amplitud observada del IC W no sea mayor que el valor deseado ω; $X^2_{gm-1,\gamma}$ es el cuantil γ de la distribución

Procedure for calculating optimal sample sizes

The CI amplitude for the mean is $2Z_{1-\alpha/2}\sqrt{s^2_p/ g}$, where. $s^2_p= (g -1)^{-1} \Sigma^g_{p=1} (m^{-1}\Sigma^m_{r=1}y_{p,r} - \bar{Y})^2$ W should be less than a specified ω value with γ probability. So, following the logic of Hahn and Meeker (1991) and Montesinos-López *et al*. (2011) for determining a sample size that ensures narrow CIs, the optimal sample size is the smallest integer value g_m so that:

$$P\{W \leq \omega\} \geq \gamma$$

$$P\{2Z_{1-\alpha/2}\sqrt{s^2_p / g} \leq \omega\} \geq \gamma$$

$$P((gm - 1)s^2_p/ \sigma^2_p \leq gm(gm - 1)\omega^2 / 4Z^2_{1-\alpha/2}\sigma^2_p)) \geq \gamma, \sigma^2_p = (k^{-1}\sigma^2 + m^{-1}\sigma^2)$$

$$\text{así,...}P(X^2_{gm-1} \leq gm\,(gm - 1)\omega^2 / 4Z^2_{1-\alpha/2}\sigma^2_p)) = \gamma$$

$$\Leftrightarrow gm\,(gm - 1)\omega^2 / 4Z^2_{1-\alpha/2}\sigma^2_p)) = X^2_{gm-1,\gamma}$$

Por lo tanto, el número de grupos requeridos es

$$gm= 4Z^2_{1-\alpha/2}(k^{-1}\sigma^2 + m^{-1}\sigma^2)\,(X^2_{gm-1,\gamma}) / (\omega^2(gm - 1) \qquad (3)$$

Where: γ is the degree of certainty (required probability) to ensure that the observed W amplitude of the CI is not higher than the desired value ω; $X^2_{gm-1,\gamma}$ is the γ quantile of the chi-square distribution with degrees of liberty. Using equation (3) the required sample size g_m is obtained, ensuring that the CI W is lower than or equal to the desired amplitude ω with the probability of at least γ. Note that if the desired level of certainty is $\gamma= 0.5$ the amount $X^2_{gm-1,\gamma}/(gm - 1)$ is approximately equal to 1, so that the equation (3) reduces to equation (2), although the equation (3) considers the randomness of the path and estimates of σ^2 and σ^2_ε by the degree of certainty desired γ. However, one drawback to derive the accurate sample size is that g_m is on both sides of the equation (3), requiring an iterative procedure to solve the equation in terms of g_m.

Approximation of the optimal sample size

When $g_m= g$ is used [Obtained in the equation (2)] on the right side of the equation (3), it has an analytical solution. This implies that g_m is equal to:

$$gm=4Z^2_{1-\alpha/2}(\sigma^2 / k + \sigma^2_\varepsilon / m)(X^2_{g-1,\gamma}) / (w^2(g - 1) = g(X^2_{g-1,\gamma}) / (g - 1) \qquad (4)$$

Where: $X^2_{g-1,\gamma}$ is the quantile γ of the chi-square distribution with $g - 1$ degrees of liberty and g is the sample size obtained with equation (2).

chi-cuadrada con g_m - 1 grados de libertad. Con la ecuación (3) se obtiene el tamaño de muestra requerido g_m, asegurando que el IC W sea menor o igual a la amplitud deseada ω con una probabilidad de al menos γ. Note que si el nivel de certeza deseado es γ= 0.5, la cantidad $X^2_{gm\text{-}1,\gamma}$ / (gm - 1) es aproximadamente igual a 1, por lo que la ecuación (3) se reduce a la ecuación (2), aunque la ecuación (3) considera la aleatoriedad de los estimadores de σ^2 y σ^2_ε vía el grado de certeza deseado γ. Sin embargo, un inconveniente para derivar el tamaño de muestra exacto es que g_m esta en ambos lados de la ecuación (3), requiriéndose de un procedimiento iterativo para resolver la ecuación en términos de g_m.

Aproximación del tamaño de muestra óptimo

Si se usa g_m= g [obtenido en la ecuación (2)] en el lado derecho de la ecuación (3), se tiene una solución analítica. Esto implica que g_m sea igual a:

$$gm=4Z^2_{1\text{-}\alpha/2}(\sigma^2/k+\sigma^2_\varepsilon/m)(X^2_{g\text{-}1,\gamma})/(w^2(g-1)=g(X^2_{g\text{-}1,\gamma})/(g-1) \quad (4)$$

Donde: $X^2_{g\text{-}1,\gamma}$ es el cuantil γ de la distribución chi-cuadrada con g - 1 grados de libertad y g es el tamaño de muestra obtenido con la ecuación (2).

Resultados y discusión

Usando el programa R (R Development Core Team, 2007), en el apéndice se proporciona información para implementar los métodos propuestos y así obtener los tamaños de muestra para cualquier combinación de σ^2, σ^2_ε, m, ω, γ y α. Los valores del Cuadro 1 fueron calculados con el método exacto [ecuación (3)]. Estos valores están basados en resultados para detectar y estimar la expresión media de un gen.

Tamaño de muestra, cuadro 1

Suponga que un investigador está interesado en estimar la expresión media de un gen de maíz GM, en la región de Oaxaca, México, donde Quist y Chapela (2001) reportaron el hallazgo de transgenes de maíz. Con esta información y después de revisar estudios previos, se hipotetiza que la varianza biológica es σ^2= 0.1, la varianza técnica es σ^2_ε= 0.02, con IC de 95%, tamaño de grupos $k=10$, réplicas técnicas igual $m=2$, y se desea que la amplitud observada del IC sea menor o igual a 0.05, es decir W_x= (μ_U - μ_L) ≤ ω= 0.05. La aplicación del método exacto señala que se requiere una

Results and discussion

Using the program R (R Development Core Team, 2007) in the appendix, provides information for implementing the proposed methods and obtain the sample sizes for any combination of σ^2, σ^2, m, ω, γ y α. The values in Table 1 were calculated with the exact method [equation (3)]. These values are based on results to detect and estimate the mean expression of a gene.

Sample size, table 1

Let's assume that a researcher is interested in estimating the mean expression of a gene of GM maize in the region of Oaxaca, Mexico, where Quist and Chapela (2001) reported finding transgenes in maize. With this information and after reviewing previous studies, it is hypothesized that the biological variance is σ^2= 0.1, the technique variance is σ^2_ε= 0.02, with 95% CI, group size k= 10, the same technical replicates m= 2, and it is desired that the observed amplitude of CI is lower than or equal to W_x= (μ_U - μ_L) ≤ ω= 0.05. The application states that the exact method requires a preliminary sample of g= 123 groups, each of size k= 10. This sample size is in the first sub-table in Table 1, with γ= 0.5, k=10, σ^2= 0.1, σ^2_ε= 0.02 y ω= 0.05.

Knowing that g= 123 produce accurate groups only 50% of the time, the researcher joins to the certainty of estimation γ= 0.99, which implies that the amplitude of CI is 95% reliable, higher than the amplitude required ω= 0.05 not more than 1% of the time. In the third sub-table in Table 1 (g_m with γ= 0.99) shows that, the sample size with the modified procedure yields g_m= 158 groups. Therefore, using 158 groups, we get a 99% certainty of the amplitude observed W of the CI will not be larger than ω= 0.05 estimating the mean μ. This sample size is located in the third sub-table in Table 1 (g_m with γ= 0.99, k=10, σ^2= 0.1, σ^2_ε= 0.02 and ω= 0.05). The use of Table 2 is quite similar, except that it contains different values for the desired amplitude (ω) and only one value for the technique variance (σ^2_ε= 0.0125).

Comparison of exact and approximate methods using groups size of k = 5.

With k= 5, equation (4) yields almost the same results obtained with the exact method [equation (3) and considering γ= 0.5]. However, if γ= 0.9 the differences in the number of groups between both methods are one or two groups;

muestra preliminar de g= 123 grupos, cada uno de tamaño k=10. Este tamaño de muestra se encuentra en el primer sub-cuadro del Cuadro 1, con γ= 0.5, k=10, σ^2= 0.1, σ_ε^2= 0.02 y ω= 0.05.

Cuadro 1. Tamaño de muestra, número de grupos, calculados con el método exacto (ecuación 3)[a].
Table 1. Sample size, number of groups, calculated with the exact method (Equation 3)[a].

	k= 5				k= 10				k= 20			
	σ_ε^2				σ_ε^2				σ_ε^2			
σ^2	0.01	0.02	0.03	0.04	0.01	0.02	0.03	0.04	0.01	0.02	0.03	0.04
		γ= 0.5				γ= 0.5				γ= 0.5		
0.05	92	123	153	184	61	92	123	153	46	77	107	138
0.1	153	184	215	246	92	123	153	184	61	92	123	153
0.15	215	246	276	307	123	153	184	215	77	107	138	169
0.2	276	307	338	369	153	184	215	246	92	123	153	184
0.25	338	369	399	430	184	215	246	276	107	138	169	200
0.3	399	430	461	492	215	246	276	307	123	153	184	215
0.35	461	492	522	553	246	276	307	338	138	169	200	230
0.4	522	553	584	614	276	307	338	369	153	184	215	246
0.45	584	614	645	676	307	338	369	399	169	200	230	261
0.5	645	676	707	737	338	369	399	430	184	215	246	276
		γ= 0.9				γ= 0.9				γ= 0.9		
0.05	109	143	176	208	75	109	143	176	58	92	126	159
0.1	176	208	241	274	109	143	176	208	75	109	143	176
0.15	241	274	306	339	143	176	208	241	92	126	159	192
0.2	306	339	371	403	176	208	241	274	109	143	176	208
0.25	371	403	435	467	208	241	274	306	126	159	192	225
0.3	435	467	499	531	241	274	306	339	143	176	208	241
0.35	499	531	563	595	274	306	339	371	159	192	225	257
0.4	563	595	627	659	306	339	371	403	176	208	241	274
0.45	627	659	691	723	339	371	403	435	192	225	257	290
0.5	691	723	754	786	371	403	435	467	208	241	274	306
		γ= 0.99				γ= 0.99				γ= 0.99		
0.05	122	158	193	228	86	122	158	193	67	104	140	176
0.1	193	228	262	296	122	158	193	228	86	122	158	193
0.15	262	296	330	363	158	193	228	262	104	140	176	210
0.2	330	363	397	430	193	228	262	296	122	158	193	228
0.25	397	430	464	497	228	262	296	330	140	176	210	245
0.3	464	497	530	563	262	296	330	363	158	193	228	262
0.35	530	563	596	629	296	330	363	397	176	210	245	279
0.4	596	629	662	695	330	363	397	430	193	228	262	296
0.45	662	695	727	760	363	397	430	464	210	245	279	313
0.5	727	760	793	825	397	430	464	497	228	262	296	330

[a] k es el tamaño de muestra; σ^2 es la varianza biológica; σ_ε^2 es la varianza técnica; IC de 95%; amplitud deseada del IC ω= 0.05; γ es el grado de certeza deseado para que el IC observado no sea más grande que la amplitud deseada ω; el número de réplicas técnicas es m= 2.

Sabiendo que los g= 123 grupos producirán ICs precisos sólo 50% de las veces, el investigador incorpora a la estimación una certeza de γ= 0.99, lo cual implica que la amplitud de

in which case the approximate method produces a slight overestimation. Also, if γ= 0.99, the approximate method produces four to six groups more that calculated with the exact method. This indicates that if γ= 0.99, the difference between both approaches increases, in such case the analytical formula lightly overestimates the optimal number of groups.

95% de confiabilidad del IC sea mayor a la amplitud requerida ω= 0.05 no más de 1% de las veces. En el tercer sub-cuadro del Cuadro 1 (g_m con γ= 0.99) se observa que el tamaño de muestra con el procedimiento modificado arroja g_m= 158 grupos. Por lo tanto, usando 158 grupos se tendrá una certeza de 99% de que la amplitud observada W del IC no será más grande que ω= 0.05 al estimar la media μ. Este tamaño de muestra se localiza en el tercer sub-cuadro del Cuadro 1 (g_m con γ=0.99, k=10, σ^2=0.1, σ^2_ε=0.02 y ω= 0.05). El uso del Cuadro 2 es similar, con la diferencia de que éste contiene diferentes valores para la amplitud deseada (ω) y sólo un valor para la varianza técnica (σ^2_ε= 0.0125).

On the other hand, using samples of size k= 20 (Table 4), if γ= 0.5 the number of groups required for both methods is the same. If γ= 0.9, the method requires approximately one to two groups more that the exact method. However, if γ= 0.99, the method requires approximately five to six groups more that exact method, indicating a slight overestimation of the optimal number of groups, just like the groups of size k= 5. But the advantage of the approximate method [equation (4)] is that it has an analytical solution, which is a simple closed formula.

Cuadro 2. Tamaño de muestra, número de grupos, calculados con el método exacto (Ecuación 3)[b].
Table 2. Sample size, number of groups, calculated with the exact method (Equation 3)[b].

	k= 5				k= 10				k= 20			
	σ^2				σ^2				σ^2			
ω	0.05	0.15	0.25	0.35	0.05	0.15	0.25	0.35	0.05	0.15	0.25	0.35
	γ= 0.50				γ= 0.50				γ= 0.50			
0.01	2497	5570	8643	11716	1728	3265	4802	6338	1344	2113	2881	3649
0.02	624	1392	2161	2929	432	816	1200	1584	336	528	720	912
0.03	277	619	960	1302	192	363	533	704	149	235	320	405
0.04	156	348	540	732	108	204	300	396	84	132	180	228
0.05	100	223	346	468	69	130	192	253	54	84	115	146
0.06	69	155	240	325	48	91	133	176	37	59	80	101
0.07	51	113	176	239	35	66	98	129	27	43	59	74
0.08	39	87	135	183	27	51	75	99	21	33	45	57
0.09	31	69	107	144	21	40	59	78	16	26	35	45
0.1	25	56	86	117	17	32	48	63	13	21	29	36
	γ= 0.90				γ= 0.90				γ= 0.90			
0.01	2587	5705	8811	11912	1803	3368	4927	6482	1410	2195	2978	3758
0.02	669	1460	2244	3027	469	867	1263	1656	369	569	768	966
0.03	307	663	1016	1367	217	397	575	752	171	262	352	441
0.04	178	381	582	781	126	229	331	432	100	152	204	255
0.05	118	249	379	507	84	151	217	282	67	101	134	167
0.06	84	177	268	358	60	107	154	200	48	72	96	119
0.07	63	133	200	267	46	81	115	149	37	55	72	90
0.08	50	103	156	207	36	64	90	117	29	43	57	70
0.09	41	83	125	166	29	51	73	94	24	35	46	57
0.1	34	69	103	136	25	43	60	77	20	29	38	47
	γ= 0.99				γ= 0.99				γ= 0.99			
0.01	2660	5814	8947	12071	1864	3451	5028	6598	1463	2262	3056	3846
0.02	705	1514	2312	3105	499	909	1313	1714	395	602	807	1010
0.03	331	699	1061	1419	236	424	608	790	188	284	377	470
0.04	196	408	615	820	141	250	356	460	113	168	223	276
0.05	131	270	405	538	95	167	236	304	77	113	149	184
0.06	96	194	290	383	70	121	170	218	56	83	108	133
0.07	73	147	219	288	54	92	129	165	44	64	83	102
0.08	58	116	172	226	43	73	102	130	35	51	66	81
0.09	48	95	139	183	36	60	83	106	29	42	54	66
0.1	40	79	116	151	30	50	70	88	25	35	45	55

[b] k es el tamaño de muestra; σ^2 es la varianza biológica; σ^2_ε es la varianza técnica (σ^2_ε = 0.0125); IC de 95%; γ es el grado de certeza deseado para que el IC observado no sea más grande que la amplitud deseada ω; el número de réplicas técnicas es m= 2.

Comparación de los métodos exacto y aproximado usando grupos de tamaño k= 5

Con k= 5, la ecuación (4) arroja casi los mismos resultados que se obtienen con el método exacto [ecuación (3) y considerando γ= 0.5]. Sin embargo, si γ= 0.9, las diferencias en el número de grupos entre los dos métodos son uno o dos grupos; en este caso el método aproximado produce una ligera sobreestimación. También, si γ= 0.99, el método aproximado produce entre cuatro y seis grupos más que los calculados con el método exacto. Esto indica que si γ= 0.99, la diferencia entre los dos enfoques se incrementa, en tal caso la fórmula analítica sobreestima ligeramente el número óptimo de grupos.

Optimal sample size-example using the formula proposed

Let´s assume that a researcher is interested in estimating the mean expression of a gene of GM plants and do not have access to Tables 2 and 3, or the package R. The researcher hypothesized that the biological variance and variance technique are σ^2= 0.1 and σ^2_ε= 0.02, respectively. Also, CI is 95% ($Z_{1-0.05/2}$= 1.96), the group size is k= 10, the technique replicas are m= 2, and it is desired that the final amplitude of CI is lower or equal to 0.05, *i.e.* W_x= (μ_U - μ_L) ≤ ω= 0.05. First, we calculate the initial sample size with the equation (2):

Cuadro 3. Comparación de los tamaños de muestra, número de grupos, de los dos métodos[c].
Table 3. Comparison of sample sizes, number of groups of two methods[c].

	Método exacto (ecuación 3)				Método analítico (ecuación 4)							
		σ^2				σ^2				Diferencia		
ω	0.05	0.15	0.25	0.35	0.05	0.15	0.25	0.35	0.05	0.15	0.25	0.35
		γ= 0.50				γ= 0.50						
0.01	2497	5570	8643	11716	2497	5570	8643	11716	0	0	0	0
0.02	624	1392	2161	2929	624	1392	2161	2929	0	0	0	0
0.03	277	619	960	1302	277	619	960	1302	0	0	0	0
0.04	156	348	540	732	156	348	540	732	0	0	0	0
0.05	100	223	346	468	100	223	346	468	0	0	0	0
0.06	69	155	240	325	69	155	240	325	0	0	0	0
0.07	51	113	176	239	51	114	176	239	0	-1	0	0
0.08	39	87	135	183	39	87	135	183	0	0	0	0
0.09	31	69	107	144	31	69	107	144	0	0	0	0
0.1	25	56	86	117	25	56	86	117	0	0	0	0
		γ= 0.90				γ= 0.90						
0.01	2587	5705	8811	11912	2588	5706	8813	11914	-1	-1	-2	-2
0.02	669	1460	2244	3027	670	1461	2246	3028	-1	-1	-2	-1
0.03	307	663	1016	1367	309	665	1017	1368	-2	-2	-1	-1
0.04	178	381	582	781	180	383	583	782	-2	-2	-1	-1
0.05	118	249	379	507	119	251	380	509	-1	-2	-1	-2
0.06	84	177	268	358	85	178	269	359	-1	-1	-1	-1
0.07	63	133	200	267	65	134	201	268	-2	-1	-1	-1
0.08	50	103	156	207	51	105	157	209	-1	-2	-1	-2
0.09	41	83	125	166	42	85	126	167	-1	-2	-1	-1
0.1	34	69	103	136	35	70	104	138	-1	-1	-1	-2
		γ= 0.99				γ= 0.99						
0.01	2660	5814	8947	12071	2665	5819	8953	12076	-5	-5	-6	-5
0.02	705	1514	2312	3105	710	1519	2317	3111	-5	-5	-5	-6
0.03	331	699	1061	1419	336	704	1066	1424	-5	-5	-5	-5
0.04	196	408	615	820	201	413	620	825	-5	-5	-5	-5
0.05	131	270	405	538	136	275	410	543	-5	-5	-5	-5
0.06	96	194	290	383	100	199	295	388	-4	-5	-5	-5
0.07	73	147	219	288	78	152	224	294	-5	-5	-5	-6
0.08	58	116	172	226	63	121	177	231	-5	-5	-5	-5
0.09	48	95	139	183	53	100	144	188	-5	-5	-5	-5
0.1	40	79	116	151	45	84	121	156	-5	-5	-5	-5

[c] IC de 95%; muestras de tamaño k= 5; σ^2_ε = 0.0125; ω es la amplitud deseada del IC; γ es el grado de certeza para que el IC observado no sea más grande que la amplitud deseada ω. La diferencia es el tamaño de muestra del método exacto menos el tamaño de muestra del método analítico.

Por otro lado, usando muestras de tamaño k= 5 (Cuadro 4), si γ= 0.5, el número de grupos requeridos con ambos métodos es el mismo. Si γ= 0.9, el método aproximado requiere entre uno y dos grupos más que el método exacto. Sin embargo, si γ= 0.99, el método aproximado necesita entre cinco y seis grupos más que el método exacto, indicando una ligera sobreestimación del número óptimo de grupos, igual a lo ocurrido con grupos de tamaño k=5. Pero la ventaja del método aproximado [ecuación (4)] es que tiene solución analítica, la cual es una fórmula cerrada muy simple.

$g= (2Z_{1-\alpha/2} / \omega)^2(\sigma^2 / k + \sigma_\varepsilon^2 / m)= (2(1.96) / 0.05)^2(0.1 / 10 + 0.02 / 2)= 123$

That is, we require g= 123 groups, which will lead to CIs with the desired precision only 50% of the time. For this reason, the researcher incorporates certain γ= 0.99, for which the equation (4) is used, thus obtaining the modified sample size that will allow to achieve their goals. As the results indicate that the preliminary sample is g= 123 and quantile γ= 0.99 of the chi-square distribution with g - 1= 122 degrees of

Cuadro 4. Comparación de los tamaños de muestra, número de grupos, de los dos métodos[d].
Table 4. Comparison of sample sizes, number of groups of two methods[d].

	Método exacto (ecuación 3) σ^2				Método analítico (ecuación 4) σ^2				Diferencia			
ω	0.05	0.15	0.25	0.35	0.05	0.15	0.25	0.35	0.05	0.15	0.25	0.35
	γ= 0.50				γ= 0.50							
0.01	1344	2113	2881	3649	1344	2113	2881	3649	0	0	0	0
0.02	336	528	720	912	336	528	720	912	0	0	0	0
0.03	149	235	320	405	149	235	320	405	0	0	0	0
0.04	84	132	180	228	84	132	180	228	0	0	0	0
0.05	54	84	115	146	54	84	115	146	0	0	0	0
0.06	37	59	80	101	37	59	80	101	0	0	0	0
0.07	27	43	59	74	27	43	59	74	0	0	0	0
0.08	21	33	45	57	21	33	45	57	0	0	0	0
0.09	16	26	35	45	16	26	35	45	0	0	0	0
0.1	13	21	29	36	13	21	29	36	0	0	0	0
	γ= 0.90				γ= 0.90							
0.01	1410	2195	2978	3758	1412	2197	2979	3760	-2	-2	-1	-2
0.02	369	569	768	966	370	571	770	968	-1	-2	-2	-2
0.03	171	262	352	441	173	263	353	443	-2	-1	-1	-2
0.04	100	152	204	255	102	154	205	256	-2	-2	-1	-1
0.05	67	101	134	167	68	102	136	169	-1	-1	-2	-2
0.06	48	72	96	119	49	74	97	121	-1	-2	-1	-2
0.07	37	55	72	90	38	56	74	91	-1	-1	-2	-1
0.08	29	43	57	70	30	44	58	72	-1	-1	-1	-2
0.09	24	35	46	57	25	36	47	58	-1	-1	-1	-1
0.1	20	29	38	47	21	30	40	48	-1	-1	-2	-1
	γ= 0.99				γ= 0.99							
0.01	1463	2262	3056	3846	1469	2267	3061	3852	-6	-5	-5	-6
0.02	395	602	807	1010	400	607	812	1015	-5	-5	-5	-5
0.03	188	284	377	470	193	289	383	475	-5	-5	-6	-5
0.04	113	168	223	276	118	173	228	281	-5	-5	-5	-5
0.05	77	113	149	184	82	118	154	189	-5	-5	-5	-5
0.06	56	83	108	133	61	88	113	138	-5	-5	-5	-5
0.07	44	64	83	102	48	68	88	107	-4	-4	-5	-5
0.08	35	51	66	81	40	56	71	86	-5	-5	-5	-5
0.09	29	42	54	66	34	47	59	71	-5	-5	-5	-5
0.1	25	35	45	55	30	40	50	60	-5	-5	-5	-5

[d] IC de 95%; muestras de tamaño k= 20; σ_ε^2 = 0.0125; ω es la amplitud deseada del IC; γ es el grado de certeza deseado para que el IC observado no sea más grande que la amplitud deseada ω. La diferencia es el tamaño de muestra del método exacto menos el tamaño de muestra del método analítico.

Tamaño de muestra óptimo- ejemplo usando la fórmula propuesta

Suponga que un investigador está interesado en estimar la expresión media de un gen de plantas GM y no tiene acceso a los Cuadros 2 y 3, ni al paquete R. El investigador hipotetiza que la varianza biológica y la varianza técnica son $\sigma^2 = 0.1$ y $\sigma_\varepsilon^2 = 0.02$, respectivamente. Además, el IC es de 95% ($Z_{1-0.05/2} = 1.96$), el tamaño del grupo es $k=10$, las réplicas técnicas son $m=2$, y se desea que la amplitud final del IC sea menor o igual a 0.05, es decir $W_x = (\mu_U - \mu_L) \leq \omega = 0.05$. Primero se calcula el tamaño de muestra inicial con la ecuación (2):

$g = (2Z_{1-\alpha/2} / \omega)^2(\sigma^2 / k + \sigma_\varepsilon^2 / m) = (2(1.96) / 0.05)^2(0.1 / 10 + 0.02 / 2) = 123$

Es decir, se requieren $g = 123$ grupos, los cuales conllevarán a ICs con la precisión deseada sólo 50% de las veces. Por esta razón, el investigador incorpora una certeza de $\gamma = 0.99$, para lo cual usa la ecuación (4), obteniendo así el tamaño de muestra modificado que le permitirá lograr sus objetivos. Como los resultados indican que la muestra preliminar es $g = 123$ y el cuantil $\gamma = 0.99$ de la distribución chi-cuadrada con $g - 1 = 122$ grados de libertad, es $X^2_{122,0.99} = 161.2495$, el tamaño de muestra modificado es $g_m = g(X^2_{g-1,\gamma}) / (g - 1) = 123(161.2495) / 122 = 163$. Esto significa que el número de grupos requeridos es 163, mientras el método exacto necesita 158, cinco grupos menos que el método aproximado. Note que la ecuación (4) produce una ligera sobreestimación pero con la ventaja de que puede determinarse fácilmente.

En general, dentro de un rango especifico para k y γ, los resultados de la fórmula fueron muy precisos, aunque la fórmula propuesta sobreestima el número óptimo de grupos, principalmente para $\gamma \geq 0.99$. Es importante señalar que la fórmula derivada asume normalidad de los datos. Por lo tanto, antes de aplicar la expresión propuesta se necesita corroborar este supuesto. Además, los métodos presentados asumen sensibilidad y especificidad perfecta, los cuales deben considerarse al diseñar el estudio.

Conclusiones

Los cuadros proporcionan los tamaños de muestra pertinentes de una amplia variedad de escenarios para estimar la expresión media de un gen, garantizando ICs precisos. Sin embargo, los resultados no cubren todas

liberty, that is $X^2_{122,0.99} = 161.2495$, the modified sample size is $g_m = g(X^2_{g-1,\gamma}) / (g - 1) = 123(161.2495) / 122 = 163$. That means that the number of groups required is 163, while the exact method requires 158, less than five groups than the approximate method. Note that equation (4) produces a slight overestimation but with the advantage that it can be easily determined.

In general, within a specific range for k and y, the results of the formula were quite accurate, although the proposed formula overestimates the optimal number of groups, mainly for $\gamma \geq 0.99$. It´s noteworthy that, the resulting formula assumes normality data. Therefore, before applying the proposed expression is needed to corroborate this assumption. Furthermore, the methods presented assume perfect sensitivity and specificity, which must be considered when designing the study.

Conclusions

The tables provide the relevant sample sizes of a wide variety of scenarios to estimate the mean expression of a gene, ensuring accurate CIs. However, the results do not cover all combinations of k, σ^2, σ_ε^2, m, ω, γ, y α; so it is recommended to use the formula for the optimal sample size [equation (3)] and the program developed in the R package (R Development Core Team, 2007) as it allows the users to determine the sample size easily and rapidly according to the requirements or needs. However, if the researcher does not have access to the program R, the practical solution is to use the equation (4). This analytical formula [equation (4)] has the advantage over the exact computational method [equation (3)] because the program R for appropriate sample sizes is needed.

Also, it is preferable to use the analytical formula than the standard method [equation (2)] since the latter yields smaller sample sizes, which produce very low probability of achieving the inferential objectives (typically lower than 0.5). We recommend using the analytical formula for group sizes smaller than or equal to 25. This recommendation is analogous to those discussed in several studies testing when used with continuous random variables groups (Kendziorski *et al.*, 2003; Dobbin and Simon, 2005; Zhang and Gant, 2005).

End of the English version

las combinaciones de k, σ^2, σ^2_ε, m, ω, γ, y α; por lo que se recomienda usar la fórmula del tamaño de muestra óptimo [ecuación (3)] y el programa desarrollado en el paquete R (R Development Core Team, 2007) ya que permite a los usuarios determinar el tamaño de muestra de manera fácil y rápida de acuerdo a los requerimientos o necesidades. No obstante, si el investigador no tiene acceso al programa R, la solución práctica es usar la ecuación (4). Esta fórmula analítica [ecuación (4)] tiene la ventaja sobre el método computacional exacto [ecuación (3)] porque no necesita del programa R para obtener tamaños de muestra apropiados.

Además, es preferible usar la fórmula analítica que el método estándar [ecuación (2)] puesto que este último arroja tamaños de muestra más pequeños, los cuales producirán probabilidades muy bajas de lograr los objetivos inferenciales (típicamente menor que 0.5). Se recomienda usar la fórmula analítica con tamaños de grupo menores o iguales a 25. Esta recomendación es análoga con las expuestas en varios estudios cuando se usan pruebas de grupo con variables aleatorias continuas (Kendziorski *et al*., 2003; Dobbin y Simon, 2005; Zhang y Gant, 2005).

Literatura citada

Beal, S. L. 1989. Sample size determination for confidence intervals on the population mean and on the difference between two population means. Biometrics. 45(3):969-977.

Caudill, S. P. 2010. Characterizing populations of individuals using pooled samples characterization. J. Exp. Sci. Environ. Epidemiol. 20(1):29-37.

Dobbin, K. and Simon, R. 2005. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics. 6(1):27-38.

Dodd, R.; Notari, E. and Stramer, S. 2002. Current prevalence and incidence of infectious disease markers and estimated window-period risk in the American Red Cross donor population. Transfusion. 42(8):975-979.

Dorfman, R. 1943. The detection of defective members of large populations. The Annals of Mathematical Statistics. 14(4):436-440.

Dyer, G. A.; Serratos-Hernández, J. A.; Perales, H. R.; Gepts, P.; Piñeyro-Nelson, A.; Chávez, A.; Salinas-Arreorta, N.; Yúnez-Naude, A.; Taylor, J. E. and Álvarez-Buylla, E. R. 2009. Dispersal of transgenes through maize seed systems in Mexico. PLoS ONE. 4(5):e5734.

Hahn, G. J. and Meeker, W. A. 1991. Statistical intervals: a guide for practitioners. Hoboken, NJ: John Wiley and Sons, Inc. 392 p.

Hernández-Suárez, C. M.; Montesinos-López, O. A.; McLaren, G. and Crossa, J. 2008. Probability models for detecting transgenic plants. Seed Sci. Res. 18(2):77-89.

Kelley, K. 2007. Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. Behavior Res. Methods. 39(4):755-766.

Kelley, K. and Maxwell, S. E. 2003. Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. Psychol. Methods. 8(3):305-321.

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

Kelley, K.; Maxwell, S. E. and Rausch, J. R. 2003. Obtaining power or obtaining precision: delineating methods of sample size planning. Eval. Health Profess. 26(3):258-287.

Kendziorski, C. M.; Zhang, Y.; Lan, H. and Attie, A. 2003. The efficiency of pooling mRNA in microarray experiments. Biostatistics. 4(3):465-477.

Kupper, L. L. and Hafner, K. B. 1989. How appropriate are popular sample size formulas? The American Statistician. 43(2):101-105.

Mary-Huard, T.; Daudin, J.; Baccinim.; Biggeri, A. and Bar-Hen A. 2007. Biases induced by pooling samples in microarray experiments. Bioinformatics. 23(13):i313-i318.

Montesinos-López, O. A.; Montesinos-López, A.; Crossa, J.; Eskridge, K. and Hernández-Suárez, C. M. 2010. Sample size for detecting and estimating the proportion of transgenic plants with narrow confidence intervals. Seed Sci. Res. 20(2):123-136.

Montesinos-López, O. A.; Montesinos-López, A.; Crossa, J.; Eskridge, K. and Sáenz, R. A. 2011. Optimal sample size for estimating the proportion of transgenic plants using the Dorfman model with a random confidence interval. Seed Sci. Res. 21(3):235-246.

Pan, Z. and Kupper, L. 1999. Sample size determination for multiple comparison studies treating confidence interval width as random. Statistics in Medicine. 18(12):1475-1488.

Peck, C. 2006. Going after BVD. Beef. 42:34-44.

Peng X.; Wood, C. L.; Blalock, E. M.; Chen, K. C.; Landfield, P. W. and Stroberg, A. J. 2003. Statistical implications of pooling RNA samples for microarray experiments. BMC Bioinformatics. 4:26.

Piñeyro-Nelson, A.; Van Heerwaarden, J.; Perales, H. R.; Serratos-Hernández, J. A. and Rangel, A. 2009. Transgenes in Mexican maize: molecular evidence and methodological considerations for GMO detection in landrace populations. Mol. Ecol. 18(4):750-761.

Quist, D. and Chapela, I. H. 2001. Transgenic DNA introgressed into traditional maize landraces in Oaxaca, Mexico. Nature. 414:541-543.

R Development Core Team. 2007. R: a language and environment for statistical computing [Computer software and manual], R Foundation for Statistical Computing. URL: http://www.r-project.org.

Remlinger, K.; Hughes-Oliver, J.; Young, S. and Lam, R. 2006. Statistical design of pools using optimal coverage and minimal collision. Technometrics. 48(1):133-143.

Schaarschmidt, F. 2007. Experimental design for one-sided confidence interval or hypothesis tests in binomial group testing. Communications in biometry and Crop Science. 2(1):32-40.

Tebbs, J. M. and Bilder, C. R. 2004. Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. J. Agric. Biol. Environ. Statistics. 9(1):75-90.

Verstraeten, T.; Farah, B.; Duchateau, L. and Matu, R. 1998. Pooling sera to reduce the cost of HIV surveillance: a feasibility study in a rural Kenyan district. Tropical Medicine and International Health. 3(9):747-750.

Wang, H.; Chow, S.C. and Chen, M. 2005. A Bayesian approach on sample size calculation for comparing means. J. Biopharmaceutical Statistics. 15(5):799-807.

Zhang, S. D. and Gant, T. W. 2005. Effect of pooling samples on the efficiency of comparative studies using microarrays. Bioinformatics. 21(24):4378-4383.