Revista Mexicana de Ciencias Agrícolas Vol.7 Núm.7 28 de septiembre - 11 de noviembre, 2016 p. 1499-1512

Tamaños de muestra que aseguran exactitud para estimar prevalencia de plantas bajo muestreo inverso*

Sample sizes that ensuring accuracy to estimate prevalence of plants under inverse sampling

Eric Eduardo Santos Fuentes[1§], Osval Antonio Montesinos López[1] y María Andrade Aréchiga[1]

[1]Universidad de Colima-Facultad de Telemática. Bernal Díaz del Castillo Núm. 340, Villas San Sebastián, 28045. Colima, México. (oamontes2@hotmail.com; mandrad@ucol.mx). [§]Autor para correspondencia: eduard77f@gmail.com.

Resumen

La detección de un evento raro o escaso (con prevalencia baja $p \leq 0.1$) en el diseño de experimentos agrícolas de una población consume muchos recursos. Por ello, se recurre al muestreo inverso (binomial negativo) el cual consta de una serie de ensayos con respuesta binaria (presencia o ausencia) en el que no se deja de muestrear hasta obtener un número predeterminado de individuos con la característica de interés. Por ello se propone un método para calcular el tamaño de muestra requerido (número de unidades positivas) bajo muestreo inverso que asegura exactitud en la proporción estimada porque garantiza que la amplitud (W) del intervalo de confianza (IC) será igual a, o más estrecha que, la amplitud deseada (ω), con una probabilidad γ (nivel de aseguramiento). Dado lo complejo y laborioso del proceso de estimación tanto del tamaño de muestra y de los parámetros de interés (proporción, varianza, desviación estándar, total e intervalos de confianza para la proporción y el total) se propone un software de distribución libre para muestreo inverso bajo el enfoque de exactitud en la estimación de parámetros que automatiza el cálculo de tamaños de muestra y de los parámetros de interés. Además el software provee una interfaz gráfica, fácil, segura y amigable con el usuario. Se recomienda el uso de la fórmula propuesta pues garantiza que con una probabilidad γ (nivel de aseguramiento ≥ 0.5) la precisión fijada a priori del IC se cumpla. Lo cual produce mayor exactitud en el estudio de interés realizado.

Abstract

The detection of a rare or scarce event (with low prevalence $p \leq 0.1$) in the design of agricultural experiments of a population consumes many resources. Therefore, one resorts to the inverse sampling (negative binomial) which consists of a series of tests with binary response (presence or absence) in which not stop sampling until a predetermined individuals with the trait of interest number. Therefore a method is proposed to calculate the required sample size (number of positive units) under inverse sampling ensures accurate estimated proportion because it ensures that the amplitude (W) of the confidence interval (IC) will be equal to, or more narrower than, the desired amplitude (ω), with a probability γ (assurance level). Given the complex and laborious process of estimating both the sample size and parameters of interest (proportion, variance, standard deviation, total and confidence intervals for the proportion and total) free software is proposed for inverse sampling under the approach of accuracy in the estimation of parameters that automates the calculation of sample sizes and parameters of interest. In addition, the software provides a graphical, easy, safe and user friendly interface. The using the formula proposed for ensuring a probability γ (assurance level ≥ 0.5) fixed a priori accuracy is met IC is recommended. Which produces more accurate study performed interest.

* Recibido: febrero de 2016
Aceptado: mayo de 2016

Palabras clave: estimación de parámetros, intervalo de confianza, prevalencia baja.

Introducción

Para garantizar la transparencia de la salud vegetal es necesario controlar la propagación de enfermedades asegurando la sanidad de las plantas en determinada población. La mejora de la sanidad vegetal tiene beneficios claros para la salud del hombre (control de enfermedades fitopatológicas, seguridad alimentaria, inocuidad y sanidad de los alimentos), repercusiones positivas para el desarrollo económico y la producción de productos agrícolas. Es por ello que los países desarrollados y en vías de desarrollo han dado mucha importancia a los programas de monitoreo y sistemas de vigilancia vegetal (Ragan, 2002).

Dentro de una población existen elementos con bajas tasas de prevalencia (un subconjunto pequeño de la población total), estos son denominados raros, escasos o evasivos (Graham y Dallas, 1988; Sudman *et al*., 1988) tienen como casos típicos plantas, animales y personas con enfermedades, tratamientos clínicos, y en general, todos aquellos grupos con baja frecuencia de unidades que poseen una característica determinada. Algunos autores como Czaja *et al.* (1996) consideran que las poblaciones raras están presentes en menos de 3% del universo de estudio.

Para estimar la prevalencia vegetal (ausencia de enfermedades en poblaciones) los métodos de muestreo son de suma importancia (Hernández-Suárez *et a*l., 2008). Por esta razón, el cálculo del tamaño de muestra óptimo es importante en el diseño de experimentos agrícolas para la estimación de proporciones en poblaciones, incluyendo la prevalencia de una enfermedad (Fosgate, 2007).

En la estimación de parámetros asociados con características escasas de una población, los diseños muestrales tradicionales no ofrecen las mejores condiciones metodológicas, principalmente por la dificultad de localizar los elementos con la característica deseada. Es por ello, que debe recurrirse a otras técnicas especiales, tales como el muestreo inverso o binomial negativo. Este método ha sido utilizado en el campo de la hematología, genética, inspección (Zhu y Lakkis, 2014), investigaciones epidemiológicas (Singh y Aggarwal, 1991; Lui, 2001; Tang *et a*l., 2008), ecología (Sheaffer y Leavenworth, 1976; Krebs, 2001), detección de

Keywords: confidence interval, low prevalence, parameter estimation.

Introduction

To ensure transparency of plant health is necessary to control the spread of ensuring the health of plants in a given population diseases. Improved plant health has clear health of man (phytopathological disease control, food safety, innocuousness and food sanitation), positive implications for economic development and agricultural production benefits. That is why the developed and developing country have given much importance to monitoring programs and monitoring of plant systems (Ragan, 2002).

Within a population there are elements with low prevalence rates (a small subset of the total population), these are called rare, rare or elusive (Graham and Dallas, 1988; Sudman *et al*., 1988) are as typical cases plants, animals and people with diseases, clinical treatments, and in general all those groups with low frequency units having a particular characteristic. Some authors like Czaja *et al.* (1996) find that rare populations are present in less than 3% of the universe of study.

To estimate the prevalence plant (absence of disease in populations) sampling methods are paramount (Hernández-Suárez *et al*., 2008). For this reason, the calculation of the optimal sample size is important in the design of agricultural experiments for estimating proportions populations, including the prevalence of disease (Fosgate, 2007).

In the estimation of parameters associated with few characteristics of a population, the traditional sample designs do not offer the best methodological conditions, mainly because of the difficulty of locating items with the desired characteristic. That is why that should be used for other special occasions, such as the inverse or negative binomial sampling techniques. This method has been used in the field of hematology, genetics, inspection (Zhu and Lakkis, 2014), epidemiological investigations (Singh and Aggarwal, 1991; Lui, 2001; Tang *et al*., 2008), ecology (Sheaffer and Leavenworth, 1976; Krebs, 2001), detection of diseases in plants and animals (Madden *et al.*, 1996), measuring the effectiveness of clinical treatments (George and Elston, 1993) among others.

enfermedades en plantas y animales (Madden *et al*., 1996), medición de la eficacia de tratamientos clínicos (George y Elston, 1993) entre otras.

Para detectar la presencia de un evento raro en una población se requiere probar un número lo suficientemente grande de individuos, y el costo de dichas pruebas por lo general excede los recursos humanos y económicos disponibles. Además de ser una actividad laboriosa, consume mucho tiempo y esfuerzo. El muestreo inverso es un método antiguo (George y Elston, 1993) para estimar una proporción p, se basa en la distribución binomial negativa con una serie de ensayos Bernoulli en el que no se deja de muestrear hasta obtener un número deseado de individuos con la característica de interés. Sin embargo, cuando la probabilidad de encontrar el atributo deseado es prácticamente nula ($p \leq 0.1$), usar el muestreo binomial (donde se fija previamente el número de elementos de la muestra) no es la mejor opción porque según Haldane (1945) el uso de una distribución binomial no siempre proporciona una estimación insesgada y precisa de p cuando ésta es pequeña ($p \leq 0.1$).

George y Elston (1993) recomiendan el uso de muestreo geométrico (que consiste en parar el proceso de muestreo hasta que se encuentra a un individuo con la característica de interés) cuando la probabilidad del evento de interés es pequeña. En su investigación se proporciona la obtención de intervalos de confianza (IC's) para la prevalencia basados en pruebas individuales y bajo un modelo geométrico. También, Haldane (1945) asevera que el uso de una distribución binomial no siempre proporciona una estimación insesgada y precisa de p cuando ésta es pequeña ($p \leq 0.1$). Lui (2000) amplió el trabajo de George y Elston (1993) para IC's al considerar el uso del muestreo binomial negativo (detener el proceso de muestreo hasta que se encuentren a $r > 1$ individuos con la característica de interés) y mostró que a medida que r aumenta, la amplitud del intervalo de confianza (IC) se reduce. Esta extensión aplica para pruebas individuales.

Históricamente, investigadores han enfatizado la planeación del tamaño de muestra en la investigación empírica, para obtener información útil de los estudios experimentales y observacionales desde una perspectiva de potencia analítica pura. Aunque la estructura de potencia analítica ha dominado la forma en que los investigadores conceptualizan la planeación del tamaño de muestra, no es ni el único, ni el mejor acercamiento que puede ser tomado para estimar el número apropiado de participantes a incluir en algún estudio de interés. Muchas veces la estimación de parámetros exactos es una meta aun potencialmente más significativa que el obtener

To detect the presence of a rare event in a population is required to try a sufficiently large number of individuals, and the cost of such tests usually exceeds human and financial resources available. Besides being a laborious activity, consuming time and effort. The inverse sampling is an ancient method (George and Elston, 1993) for estimating a proportion p, it is based on the negative binomial distribution with a series of trials Bernoulli where it is not left to sample until a desired number of individuals with characteristic of interest. However, when the probability of finding the desired attribute is practically nil ($p \leq 0.1$), using the binomial sampling (where the number of elements in the sample is preset) is not the best option because according Haldane (1945) use a binomial distribution does not always provide an unbiased and accurate estimate of p when it is small ($p \leq 0.1$).

George and Elston (1993) recommend the use of geometric sampling (consisting stop the sampling process until it is an individual with the characteristic of interest) when the probability of the event of interest is small. In his research obtaining confidence intervals (IC's) for prevalence based on individual tests and under a geometric model is provided. Also, Haldane (1945) asserts that the use of a binomial distribution does not always provide an unbiased and accurate estimate of p when it is small ($p \leq 0.1$). Lui (2000) extended the work of George and Elston (1993) for IC's to consider using negative binomial sampling (stop the sampling process until $r > 1$ individuals with the characteristic of interest are) and showed that as r increases, the amplitude of the confidence interval (IC) is reduced. This extension applies to individual tests.

Historically, researchers have emphasized planning sample size in empirical research, to obtain useful information from experimental and observational studies from a perspective of pure analytical power. Although the structure of analytical power has dominated the way researchers conceptualize planning sample size, is neither the only, nor the best approach that can be taken to estimate the appropriate number of participants to be included in any study of interest. Often estimating exact parameters is potentially even more significant goal the obtaining statistical significance (Kelley *et al*., 2003). This shows that the appropriate method for planning the sample size, and the appropriate size of the sample itself, depend on the desired goals in an investigation.

An alternative approach by Kelley (2007) for the frame of analytical power for determining sample sizes is ensuring the accuracy in estimating parameters (AIPE). The aim of

significancia estadística (Kelley *et al.,* 2003). Esto muestra que el método apropiado para la planeación del tamaño de muestra, y el tamaño apropiado de la muestra en sí, dependen de las metas deseadas en una investigación.

Un enfoque alternativo según Kelley (2007) para el marco de potencia analítica para la determinación de tamaños de muestra es el que garantiza la exactitud en la estimación de parámetros (AIPE por sus siglas en inglés). El objetivo de AIPE es garantizar que los parámetros estimados correspondan con la exactitud fijada para estimar dicho parámetro poblacional. Autores como Montesinos-López *et al.* (2011) han desarrollado procedimientos para el cálculo de tamaños de la muestra bajo el enfoque AIPE, enfoque que garantiza cortos IC's para la estimación de parámetros bajo muestreo binomial y pruebas de grupo. Los IC's también transmiten información para determinar con precisión la magnitud del efecto a partir de los datos disponibles (Beal, 1989; Montesinos-López *et al.,* 2012). Por ello, la determinación de tamaños de muestra bajo este enfoque puede contribuir a inferir en teorías más fuertes y precisas sobre algún fenómeno en estudio.

Realizar los cálculos necesarios para determinar tamaños de muestra y estimar otros parámetros importantes bajo muestreo inverso, resulta sin duda alguna una tarea laboriosa. Montesinos-López *et al.* (2012) proponen un algoritmo computacional en el paquete estadístico R, donde se calculan tamaños de muestra con enfoque AIPE para el muestreo inverso, pero cabe destacar que sólo se enfoca al muestreo por grupos (en inglés Group Testing), además de carecer de una interfaz gráfica, amigable y de fácil uso.

Por lo tanto, los propósitos de este artículo son: 1) derivar una expresión para calcular tamaños de muestra para la estimación de una proporción (p) bajo muestreo inverso (binomial negativo) bajo el enfoque AIPE, 2) mostrar a través de ejemplos el proceso de cálculo para el tamaño de muestra y 3) desarrollar un software para realizar la determinación de tamaños de muestra así como la estimación de parámetros de interés bajo este enfoque.

Materiales y métodos

Suponga que $Y_i = y_i$ individuos son probados hasta encontrar el primer individuo positivo y $Y_1, Y_2, Y_3, \ldots Y_r$ son observados para obtener el *r-ésimo* individuo positivo. Dado que, Y_i (i=

AIPE is to ensure that the estimated parameters correspond to the accuracy set to estimate that the population parameter. Authors like Montesinos-López *et al*. (2011) have developed procedures for calculating sample sizes under the AIPE approach, an approach that guarantees short IC's for estimating parameters under binomial sampling and testing group. The IC's also transmit information to accurately determine the magnitude of the effect from the available data (Beal, 1989; Montesinos-López *et al*., 2012). Therefore, the determination of sample sizes under this approach can contribute to infer stronger and precise theories about a phenomenon under study.

Perform the necessary calculations to determine sample sizes and estimate other important parameters under reverse sampling is undoubtedly a laborious task. Montesinos-López *et al.* (2012) propose a computational algorithm in the R statistical package where sample sizes are calculated focusing AIPE for reverse sampling, but it is noteworthy that only focuses on the sampling groups (Group Testing), besides lacking a graphics, friendly and easy user interface.

Therefore, the purposes of this article are: 1) to derive an expression for calculating sample sizes for estimating a proportion (p) under inverse sampling (negative binomial) under the AIPE approach, 2) show through examples calculation process for sample size and 3) develop software for the determination of sample sizes and the estimation of parameters of interest under this approach.

Materials and methods

Suppose $Y_i = y_i$ individuals are tested to find the first positive individual and $Y_1, Y_2, Y_3, \ldots Y_r$ are observed for the r-th positive individual. Since, Y_i (i= 1, 2,…,r) has a geometric distribution. Therefore, the total number of individuals is recorded to find positive individuals r is equal to $T = \sum_{i=1}^{r} Y_i$. The prevalence is denoted by *p*, the number of tested to find the first positive individuals individual is $Y_i = y_i$, and the number of times the experiment is carried out is denoted by *r*. It is important to note that this document is considered that: (i) the sample size is the value of *r* representing the required number of positive individuals to stop the process of sampling and testing, and (ii) the total number of individuals tested is the value of $T = \sum_{i=1}^{r} Y_i$. Therefore, sufficient and complete statistic $T = \sum_{i=1}^{r} Y_i$ has a negative binomial distribution (*dbn*) with *r* parameter and probability

1, 2,…,r) tiene una distribución geométrica. Por lo tanto, se registra el número total de individuos para encontrar r individuos positivos que es igual a $T=\Sigma_{i=1}^{r} Y_i$. La prevalencia es denotada por p, el número de individuos probados hasta encontrar el primer individuo positivo es $Y_i=y_i$, y el número de veces que el experimento se lleva a cabo está denotado por r. Es importante mencionar que en este documento se considera que: (i) el tamaño de muestra es el valor de r que representa el número requerido de individuos positivos para detener el proceso de muestreo y las pruebas, y (ii) el número total de individuos probados es el valor de $T=\Sigma_{i=1}^{r} Y_i$. Por lo tanto, la estadística suficiente y completa $T=\Sigma_{i=1}^{r} Y_i$ tiene una distribución binomial negativa (*dbn*) con parámetro r y probabilidad de éxito p (George y Elston, 1993). De acuerdo con George y Elston (1993) la estimación de máxima verosimilitud (*EMV*) de p usando muestreo inverso es:

$$\hat{p}=\frac{r}{T} \qquad 1)$$

Donde: r es el número fijado requerido de individuos positivos. Este EMV de p para muestreo inverso asume una prueba diagnóstica perfecta (especificidad y sensibilidad iguales a uno). Por otro lado, la varianza de p̂ de acuerdo a George y Elston (1993) está dada por $V(\hat{p})=\frac{p^2(1-p)}{r}$ y tomando en cuenta el factor de corrección de población finita es igual a $V(\hat{p})=\left[\frac{N-n}{N}\right]\left[\frac{p^2q}{r}\right]$, donde q= (1 - p). De acuerdo a George y Elston (1993) el IC de Wald es el siguiente:

$$p_L=\hat{p}-Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})}$$
$$p_U=\hat{p}-Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})} \qquad 2)$$

Donde: $Z_{1-\alpha/2}$ es el cuantil $1-\alpha/2$ de la distribución normal estándar, y p̂ es el *EMV* . Esta aproximación del IC es fácil de calcular y permite derivar fórmulas del tamaño de la muestra de forma cerrada. Sin embargo, cuando r es pequeña, la aproximación normal para *EMV* es dudosa; puesto que en tales casos, el IC de Wald frecuentemente produce puntos finales negativos. Además, la probabilidad de cobertura de los IC's construidos por el IC de Wald es frecuentemente menor que $100(1-\alpha)\%$.

La cantidad $Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})}$ (añadida y sustraída de la proporción observada, p̂) en la Ec. 2 se define como W/2 (donde W es la amplitud del IC; W o W/2 se puede establecer a priori por el investigador dependiendo de la precisión deseada). La

of success *p* (George and Elston, 1993). According to George and Elston (1993) maximum likelihood estimation (EMV) of p using inverse sampling is:

$$\hat{p}=\frac{r}{T} \qquad 1)$$

Where: *r* is the required number of fixed positive individuals. This EMV of p for reverse sampling assumes a perfect diagnostic test (specificity and sensitivity equal to one). Moreover, the variance of p̂ according to George and Elston (1993) is it given by $V(\hat{p})=\frac{p^2(1-p)}{r}$ taking into account the correction factor of finite population is equal to $V(\hat{p})=\left[\frac{N-n}{N}\right]\left[\frac{p^2q}{r}\right]$ where q= (1 - p). According to George and Elston (1993) the IC of Wald is:

$$p_L=\hat{p}-Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})}$$
$$p_U=\hat{p}-Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})} \qquad 2)$$

Where: $Z_{1-\alpha/2}$ is the quantile $1-\alpha/2$ standard normal distribution, and p̂ is the *EMV*. This approach IC is easy to calculate and derive formulas allows the size of the sample so closed. However, when *r* is small, the normal approach to *EMV* is doubtful; since in such cases the IC of Wald often produces negative endpoints. Furthermore, the probability of coverage of IC's constructed by IC of Wald is often less than $100(1-\alpha)\%$.

The amount $Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})}$ (added and subtracted from the observed ratio, p̂) in Ec. 2 is defined as W/2 (where W is the width of IC; W o W/2 can be set a priori by the researcher depending on the desired accuracy). The IC amplitude observed for any embodiment of this confidence interval (From Ec. 2) can be expressed as:

$$W=2Z_{1-\alpha/2}\sqrt{\frac{\hat{p}^2(1-\hat{p})}{r}} \qquad 3)$$

Be the desired amplitude ω IC; then the AIPE approach basically find the minimum sample size that ensures that the expected amplitude of the IC is sufficiently narrow (Kelley, 2007; Kelley and Rausch, 2011). In other words, the approach seeks AIPE minimum sample size such that $E(W) \leq \omega$. The problem is that the expected amplitude of the IC is an unknown quantity, but can be approximated.

amplitud del IC observado para cualquier realización de este intervalo de confianza (A partir de la Ec. 2) se puede expresar como:

$$W = 2Z_{1-\alpha/2}\sqrt{\frac{\hat{p}^2(1-\hat{p})}{r}} \quad 3)$$

Sea ω la amplitud deseada del IC; entonces el enfoque AIPE básicamente trata de encontrar el tamaño mínimo de la muestra que garantiza que la amplitud esperada del IC sea lo suficientemente estrecha (Kelley, 2007; Kelley y Rausch, 2011). En otras palabras, el enfoque AIPE busca el tamaño de muestra mínimo de tal manera que $E(W) \leq \omega$. El problema es que la amplitud esperada del IC es una cantidad desconocida, aunque se puede aproximar.

Por consiguiente, el valor esperado de W es:

$$E(W) = E\left[2Z_{1-\alpha/2}\sqrt{\frac{\hat{p}^2(1-\hat{p})}{r}}\right] \approx 2Z_{1-\frac{\alpha}{2}}\frac{p^2(1-p)}{r} \text{ (Lui, 1995).}$$

Ahora bien, si se establece el E(W) para la amplitud deseada del IC, ω:

$$\omega = 2Z_{1-\alpha/2}\sqrt{\frac{p^2(1-p)}{r}} \quad 4)$$

De acuerdo con Lui (2001) resolviendo r (Ec. 4) se obtiene la siguiente fórmula:

$$r_p = \frac{4Z^2_{1-\alpha/2}\ p^2(1-p)}{\omega^2} \quad 5)$$

Sin embargo, la Ec. 5 requiere el valor poblacional de p, que es desconocido y en la práctica se sustituye por una estimación de la proporción real. Aunque, la Ec. 5 se proporciona el tamaño de muestra necesario para alcanzar la amplitud deseada del IC, E(W), que es suficientemente estrecha para la estimación de la proporción. Sin embargo, esto no garantiza que para cualquier IC en particular, la amplitud esperada del IC observado, E(W), sea lo suficientemente estrecha, porque el valor esperado sólo se aproxima a la amplitud del intervalo de confianza promedio. Kelley y Rausch (2011) afirman que este problema es similar al caso donde se estima un promedio a partir de una distribución normal, aunque la media muestral es un estimador insesgado de la media poblacional, es casi seguro que la media muestral sea más pequeña o más grande que el valor de la población. Esto se debe a que la media muestral es una variable aleatoria continua, como lo es la anchura del IC, debido a que ambos se basan en datos aleatorios. Por lo tanto, aproximadamente la mitad de las veces, la anchura calculada del IC será mayor que la anchura previamente especificada (Kelley y Rausch, 2011).

Therefore, the expected value of W is:

$$E(W) = E\left[2Z_{1-\alpha/2}\sqrt{\frac{\hat{p}^2(1-\hat{p})}{r}}\right] \approx 2Z_{1-\frac{\alpha}{2}}\frac{p^2(1-p)}{r} \text{ (Lui, 1995).}$$

Now if the $E(W)$ is set to the desired amplitude of the IC, ω:

$$\omega = 2Z_{1-\alpha/2}\sqrt{\frac{p^2(1-p)}{r}} \quad 4)$$

According to Lui (2001) by solving *r* (Ec. 4) the following formula is obtained:

$$r_p = \frac{4Z^2_{1-\alpha/2}\ p^2(1-p)}{\omega^2} \quad 5)$$

However, Ec. 5 requires the population value of p, which is known in practice and is replaced by an estimate of the actual ratio. Although Ec. 5the sample size necessary to achieve the desired amplitude of the IC, E(W), which is close enough to the estimated proportion is provided. However, this does not guarantee that any particular IC, the expected amplitude of the observed IC, E(W), is sufficiently narrow, because the expected value only approximates the average interval width confidence. Kelley and Rausch (2011) argue that this problem is similar to the case where an average is estimated from a normal distribution, although the sample mean is an unbiased estimator of the population mean, it is almost certain that the sample mean is smaller or larger than the value of the population. This is because the sample mean is a continuous random variable, as is the width of the IC, because both are based on random data. Therefore, about half of the time width calculated IC is greater than the previously especified width (Kelley and Rausch, 2011).

Because Ec. 3 uses an estimate of *p*, the width of the IC (W) is a random variable fluctuate from sample to sample. This implies that r_p using Ec. 5, about 50% of the sampling distribution of W is less than ω (third column of Table 1). To demonstrate this, you calculate the probability of an amplitude of less IC to the specified value (ω). This can be calculated by:

$$P(W \leq \omega) = \sum_{t=r_p}^{\infty} I(w_t, y, p)\begin{bmatrix} t-1 \\ r_p-1 \end{bmatrix}[p]^{r_p}[1-p]^{t-r_p}$$

Where: $I(w_t, y, p)$ is an indicator function that shows whether the actual IC width calculated using Ec. 3 es ≤ ω, p is the true proportion of the population and r_p is the size of the sample obtained. Ec. 5. Due to computational limitations the following approximation to calculate this probability is used.

$$P(W \leq \omega) = \sum_{t=r_p}^{t^*} I(w_t, y, p)\begin{bmatrix} t-1 \\ r_p-1 \end{bmatrix}[p]^{r_p}[p]^{t-r_p} \quad 6)$$

Debido a que la Ec. 3 utiliza una estimación de p, la anchura del IC (W) es una variable aleatoria que fluctuará de muestra a muestra. Esto implica que usando r_p de la Ec. 5, alrededor del 50% de la distribución de muestreo de W será menor que ω (Tercer columna del Cuadro 1). Para demostrar esto, se debe calcular la probabilidad de obtener una amplitud del IC menor al valor especificado (ω). Esto se puede calcular por medio de:

$$P(W \leq \omega)= \sum_{t=r_p}^{\infty} I(w_t, y, p) \binom{t-1}{r_p-1} [p]^{r_p}[1-p]^{t-r_p}$$

Donde: I (w_t, y, p) es una función indicadora que muestra si la anchura calculada de IC real usando la Ec. 3 es ≤ ω, p es la proporción verdadera de la población y r_p es el tamaño de la muestra obtenida mediante la Ec. 5. Debido a las limitaciones computacionales se utiliza la siguiente aproximación para calcular esta probabilidad.

$$P(W \leq \omega)= \sum_{t=r_p}^{t^*} I(w_t, y, p) \binom{t-1}{r_p-1} [p]^{r_p}[p]^{t-r_p} \qquad 6)$$

Donde: $t=r_p, r_p+1, r_p+2,\ldots t^*$, y W se considera una variable aleatoria ya que el valor exacto de p no se conoce y t* es el valor que satisface $P(T \leq t^*)= 0.9999$. Se utiliza el valor de t* ya que la operación en el Paquete R no puede sumar hasta infinito.

Derivación del tamaño de muestra bajo AIPE para muestreo inverso

A continuación se muestra el procedimiento para derivar la expresión para el cálculo de tamaños de muestra para la estimación de una proporción (p) para muestreo inverso bajo el enfoque AIPE.

Where: $t=r_p, r_p+1, r_p+2,\ldots t^*$, and W is considered a random variable since the exact value *p* is not known t* is the value that satisfies $P(T \leq t^*)= 0.9999$. The value of t* it is used as the operation in the R package may not add up to infinity.

Derivation low sample size AIPE for reverse sampling

The procedure for deriving the expression for calculating sample size for estimation of a ratio (*p*) for sampling under AIPE reverse approach is shown.

Suppose $Y_1,\ldots,Y_r$ is a random sample of size *r* of a geometric distribution (p). Let $h(x)=\sqrt{(1-1/\overline{T}_r)(1/\overline{T}_r)^2}$ with $\overline{T}_r= \frac{\sum_{i=1}^{r} y_i}{r}$.

Then for $p \neq \frac{2}{3}$, $\frac{\sqrt{r}\left[h(\overline{T}_r) - h\left[\frac{1}{p}\right]\right]d}{\sqrt{h'\left[\frac{1}{p}\right]^2 \sigma_r^2}} \rightarrow N(0,1)$.

That is, $h(\overline{T}_r) = \sqrt{\left[1-\frac{1}{\overline{T}_r}\right]\left[\frac{1}{\overline{T}_r}\right]^2} \sim N\left[h\left[\frac{1}{p}\right], h'\left[\frac{1}{p}\right]^2 \sigma_r^2\right]$; where $\sigma_r^2=\frac{1-p}{rp^2}$, $h\left[\frac{1}{p}\right]=\sqrt{p^2(1-p)}$ and $h'\left[\frac{1}{p}\right]= \frac{1}{\sqrt{(1-p)p^2}}\left[\frac{3}{2}p^4 - p^3\right]$.

Note that $\overline{T}_r \sim N\left[\frac{1}{p}, \sigma_r^2=\frac{1-p}{rp^2}\right]$. Then if $\sigma_r^2 \rightarrow 0$, si $r \rightarrow \infty$, $h(x)= \sqrt{(1-1/x)(1/x)^2}$ is differentiable with respect to a $x \in (0,1)$ y $h'\left[\frac{1}{p}\right]= \frac{1}{\sqrt{(1-p)p^2}}\left[\frac{3}{2}p^4 - p^3\right] \neq 0$. For $p \frac{2}{3}$, then using the delta method (Oehlert, 1992) is obtained, $h(\overline{T}_r) \sim N\left[h\left[\frac{1}{p}\right], \left[h'\left[\frac{1}{p}\right]\right]^2 \sigma_r^2\right]$.

Cuadro 1. Resultados de la subestimación del tamaño de muestra utilizando la Ec. 5.
Table 1. Results of underestimating the size of sample using Ec. 5.

p	r_p	$P(W \leq \omega)$	r_{m10}	$P(W \leq \omega)$	r_{m20}	$P(W \leq \omega)$	r_{m40}	$P(W \leq \omega)$
0.005	8	0.4794223	18	0.9456369	28	0.9986697	48	0.999919
0.0075	17	0.4739768	27	0.8663093	37	0.9856035	57	0.9999083
0.01	31	0.4800041	41	0.7989407	51	0.9537533	71	0.999383
0.0125	48	0.4738842	58	0.7431816	68	0.9114708	88	0.9962598
0.015	70	0.4993138	80	0.7259672	90	0.8829195	110	0.9900424
0.0175	94	0.4810978	104	0.6824672	114	0.8378104	134	0.9763156
0.02	123	0.4912225	133	0.6682736	143	0.8126911	163	0.9618497
0.0225	155	0.4873608	165	0.6473447	175	0.7830901	195	0.9429291
0.025	191	0.4892276	201	0.6341471	211	0.7610802	231	0.9244332

Suponga que $Y_1,\ldots,Y_r$ es una muestra aleatoria de tamaño r de una distribución geométrica (p). Sea $h(x)=\sqrt{(1-1/\overline{T}_r)(1/\overline{T}_r)^2}$ con $\overline{T}_r=\frac{\sum_{i=1}^{r}y_i}{r}$. Entonces para $p\neq\frac{2}{3}$, $\frac{\sqrt{r}\left[h(\overline{T}_r)-h\left[\frac{1}{p}\right]\right]d}{\sqrt{h'\left[\frac{1}{p}\right]^2\sigma_r^2}}\rightarrow N(0,1)$.

Esto es, $h(\overline{T}_r)=\sqrt{\left(1-\frac{1}{\overline{T}_r}\right)\left(\frac{1}{\overline{T}_r}\right)^2}\sim N\left[h\left[\frac{1}{p}\right],h'\left[\frac{1}{p}\right]^2\sigma_r^2\right]$; donde $\sigma_r^2=\frac{1-p}{rp^2}$, $h\left[\frac{1}{p}\right]=\sqrt{p^2(1-p)}$ y $h'\left[\frac{1}{p}\right]=\frac{1}{\sqrt{(1-p)p^2}}\left[\frac{3}{2}p^4-p^3\right]$. Note que $\overline{T}_r\sim N\left[\frac{1}{p},\sigma_r^2=\frac{1-p}{rp^2}\right]$. Entonces, si $\sigma_r^2\rightarrow 0$, si $r\rightarrow\infty$, $h(x)=\sqrt{(1-1/x)(1/x)^2}$ es diferenciable con respecto a $x\in(0,1)$ y $h'\left[\frac{1}{p}\right]=\frac{1}{\sqrt{(1-p)p^2}}\left[\frac{3}{2}p^4-p^3\right]\neq 0$. Para $p\neq\frac{2}{3}$, entonces usando el método delta (Oehlert, 1992) se obtiene, $h(\overline{T}_r)\sim N\left[h\left[\frac{1}{p}\right],\left[h'\left[\frac{1}{p}\right]\right]^2\sigma_r^2\right]$.

Entonces, el valor entero más pequeño r_m tal que:

$$P\left[2Z_{1-\alpha/2}\sqrt{\frac{(1-\hat{p})(\hat{p})^2}{r_m}}\leq\omega\right]\geq\gamma$$

$$P\left[2Z_{1-\alpha/2}\sqrt{\frac{(1-\hat{p})(\hat{p})^2}{r_m}}\leq\omega\right]$$

$$=P\left[\frac{h\left[\frac{1}{\hat{p}}\right]-h\left[\frac{1}{p}\right]}{\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{r_mp^2}}}\leq\frac{\frac{\omega\sqrt{r_m}}{2Z_{1-\alpha/2}}-h\left[\frac{1}{p}\right]}{\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{r_mp^2}}}\right]=\gamma$$

$$\Leftrightarrow P\left[Z\leq\frac{\frac{\omega\sqrt{r_m}}{2Z_{1-\alpha/2}}-h\left[\frac{1}{p}\right]}{\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{r_mp^2}}}\right]\approx\gamma\Leftrightarrow\frac{\frac{\omega\sqrt{r_m}}{2Z_{1-\alpha/2}}-h\left[\frac{1}{p}\right]}{\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{r_mp^2}}}\approx Z_\gamma$$

$$\Leftrightarrow\frac{\omega}{2Z_{1-\alpha/2}}r_m-h\left[\frac{1}{p}\right]\sqrt{r_m}-Z_\gamma\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{p^2}}\approx 0$$

$$\Leftrightarrow\frac{\omega}{2Z_{1-\alpha/2}}r_m-h\left[\frac{1}{p}\right]\sqrt{r_m}-Z_\gamma\left|h'\left[\frac{1}{p}\right]\right|\sqrt{\frac{(1-p)}{p^2}}\approx 0 \qquad 7)$$

Then, the smallest integer value r_m such that:

$$P\left[2Z_{1-\alpha/2}\sqrt{\frac{(1-\hat{p})(\hat{p})^2}{r_m}}\leq\omega\right]\geq\gamma$$

$$P\left[2Z_{1-\alpha/2}\sqrt{\frac{(1-\hat{p})(\hat{p})^2}{r_m}}\leq\omega\right]$$

$$=P\left[\frac{h\left[\frac{1}{\hat{p}}\right]-h\left[\frac{1}{p}\right]}{\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{r_mp^2}}}\leq\frac{\frac{\omega\sqrt{r_m}}{2Z_{1-\alpha/2}}-h\left[\frac{1}{p}\right]}{\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{r_mp^2}}}\right]=\gamma$$

$$\Leftrightarrow P\left[Z\leq\frac{\frac{\omega\sqrt{r_m}}{2Z_{1-\alpha/2}}-h\left[\frac{1}{p}\right]}{\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{r_mp^2}}}\right]\approx\gamma\Leftrightarrow\frac{\frac{\omega\sqrt{r_m}}{2Z_{1-\alpha/2}}-h\left[\frac{1}{p}\right]}{\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{r_mp^2}}}\approx Z_\gamma$$

$$\Leftrightarrow\frac{\omega}{2Z_{1-\alpha/2}}r_m-h\left[\frac{1}{p}\right]\sqrt{r_m}-Z_\gamma\sqrt{\left[h'\left[\frac{1}{p}\right]\right]^2\frac{(1-p)}{p^2}}\approx 0$$

$$\Leftrightarrow\frac{\omega}{2Z_{1-\alpha/2}}r_m-h\left[\frac{1}{p}\right]\sqrt{r_m}-Z_\gamma\left|h'\left[\frac{1}{p}\right]\right|\sqrt{\frac{(1-p)}{p^2}}\approx 0 \qquad 7)$$

Note that Eq. 7 has a quadratic form $ax^2+bx+c=0$, with $x=\sqrt{r_m}$, $a=\frac{\omega}{2Z_{1-\alpha/2}}$, $b=-h\left[\frac{1}{p}\right]$, and $c=-Z_\gamma\left|h'\left[\frac{1}{p}\right]\right|\sqrt{\frac{(1-p)}{p^2}}$, with two solutions given by $x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$. Taking $x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$. Therefore, for a fixed ω, the r_m desired is given by $r_m=\left(\frac{h\left[\frac{1}{p}\right]+\sqrt{h\left[\frac{1}{p}\right]^2+\frac{2\omega}{Z_{1-\alpha/2}}Z_\gamma\left|h'\left[\frac{1}{p}\right]\right|\sqrt{\frac{1-p}{p^2}}}}{\frac{\omega}{Z_{1-\alpha/2}}}\right)^2$, where after replacing $h\left[\frac{1}{p}\right]$ and $h'\left[\frac{1}{p}\right]$ is obtained

$$r_m=\left(\frac{\sqrt{(1-p)p^2}+\sqrt{(1-p)p^2+\frac{2\omega}{Z_{1-\alpha/2}}Z_\gamma\left|\frac{1}{\sqrt{(1-p)p^2}}(1.5p^4-p^3)\right|\sqrt{\frac{1-p}{p^2}}}}{\frac{\omega}{Z_{1-\alpha/2}}}\right)^2$$

$$r_m=\left[\frac{Z_{1-\alpha/2}}{\omega}\right]^2\left[\sqrt{p^2(1-p)}+\sqrt{p^2(1-p)+\frac{2\omega|(1.5p^4-p^3)|Z_\gamma}{Z_{1-\alpha/2}\,p^2}}\right]^2. \qquad 8)$$

Note que la Ec. 7 tiene una forma cuadrática $ax^2 + bx + c = 0$, con $x = \sqrt{r_m}$, $a = \frac{\omega}{2Z_{1-\alpha/2}}$, $b = -h\left[\frac{1}{p}\right]$, y $c = -Z_\gamma \left|h'\left[\frac{1}{p}\right]\right| \sqrt{\frac{(1-p)}{p^2}}$, con dos soluciones dadas por $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$. Tomando $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$. Por lo tanto, para un ω fijo, el r_m deseado está dado por

$$r_m = \left(\frac{h\left[\frac{1}{p}\right] + \sqrt{h\left[\frac{1}{p}\right]^2 + \frac{2\omega}{Z_{1-\alpha/2}} Z_\gamma \left|h'\left[\frac{1}{p}\right]\right| \sqrt{\frac{1-p}{p^2}}}}{\frac{\omega}{Z_{1-\alpha/2}}} \right)^2,$$

en el que después de remplazar $h\left[\frac{1}{p}\right]$ y $h'\left[\frac{1}{p}\right]$ se obtiene

$$r_m = \left(\frac{\sqrt{(1-p)p^2} + \sqrt{(1-p)p^2 + \frac{2\omega}{Z_{1-\alpha/2}} Z_\gamma \left| \frac{1}{\sqrt{(1-p)p^2}} (1.5p^4 - p^3) \right| \sqrt{\frac{1-p}{p^2}}}}{\frac{\omega}{Z_{1-\alpha/2}}} \right)^2$$

$$r_m = \left[\frac{Z_{1-\alpha/2}}{\omega} \right]^2 \left[\sqrt{p^2(1-p)} + \sqrt{p^2(1-p) + \frac{2\omega |(1.5p^4 - p^3)| Z_\gamma}{Z_{1-\alpha/2}\, p^2}} \right]^2. \quad 8)$$

Este procedimiento muestra la expresión para el cálculo de tamaños de muestra que garantizan exactitud en la estimación de parámetros de interés.

Resultados y discusiones

Subestimación del tamaño de muestra utilizando la Ec. 5.

Para mostrar el grado en que r_p es subestimado por la Ec. 5, se proporciona un ejemplo (Cuadro 1) en el que se utiliza la Ec. 6 para calcular $P(W \leq \omega)$, es decir, la probabilidad de que W sea menor o igual a la amplitud deseada (ω) del IC para un valor dado r_p (número unidades positivas) obtenido con la Ec. 5. El ejemplo numérico en el Cuadro 1 se da para varios valores de la proporción poblacional (p) con un IC de 95%, y para una anchura deseada ω= 0.007. El Cuadro 1 presenta el tamaño de muestra preliminar calculando r_p con la Ec. 5, y otros tres incrementos: $r_{m10} = r_p + 10$, $r_{m20} = r_p + 20$ y $r_{m40} = r_p + 40$. Para cada tamaño de muestra, la probabilidad de que W sea menor que el valor especificado (ω= 0.007) es $P(W \leq \omega)$ y se calcula utilizando la Ec. 6. Esto se hace para mostrar que el número requerido de unidades positivas para estimar la proporción (r_p de la segunda columna) calculada utilizando la Ec. 5 tiene una probabilidad de alrededor de 0.50

This procedure shows the expression for calculating sample sizes ensuring accuracy in the estimation of parameters of interest.

Results and discussions

Understatement sample size using Ec. 5.

To show the extent to which r_p is underestimated by Ec. 5, an example (Table 1) in which Ec. 6 is used is provided to calculate $P(W \leq \omega)$ i.e., the probability that W is less than or equal to the desired amplitude (ω) for IC a given r_p (number positive units) obtained with Ec. 5. The numerical value example in Table 1 is given for various values of the population proportion (p) with IC at 95%, and a desired width ω= 0.007. In the Table 1 presents the preliminary sample size r_p calculating Ec 5, and three other increases: $r_{m10} = r_p + 10$, $r_{m20} = r_p + 20$ and $r_{m40} = r_p + 40$. For each sample size, the probability that *W* is less than the specified value (ω= 0.007) is $P(W \leq \omega)$ and is calculated using Ec. 6. This is done to show that the required number of positive units to estimate the proportion (r_p of the second column) calculated using Ec. 5 has a probability of about 0.50 that $W \leq \omega = 0.007$ (third column). For example, when p= 0.02, the preliminary sample size (r_p) is 48 and the probability of obtaining a $W \leq \omega = 0.007$ is 0.4738842. With $p = 0.02$, $r_p = 123$, can only be 49.12% sure that W will ≤ ω= 0.007. When the number of positive units increases by 10 (r_{m10}, fourth column) or 20 (r_{m20}, sixth column), the probability $P(W \leq \omega = 0.007)$ increases.

For example, when $p = 0.0125$, there $r_{m20} = 68$ units in the sample with $P(W \leq \omega = 0.007) = 0.9114708$; for $r_{m40} = 88$ positive units in the sample, $P(W \leq \omega = 0.007) = 0.9962598$. Therefore, the results in Table 1 show that a sample size (number of positive units) is required to ensure high $P(W \leq \omega = 0.007)$, greater than the value r_p calculated Eq. 5. Furthermore, Table 1 shows that in the 9 cases the size of the resulting preliminary sample (number of positive observations) using Ec. 5 produces at $P(W \leq \omega) < 0.50$, that is, 100% of the time $P(W \leq \omega = 0.007)$ is less than 50%.

For estimation of plant prevalence low inverse sampling

As an example of estimation for the use of the techniques mentioned develops the following case:

de que $W \leq \omega = 0.007$ (tercera columna). Por ejemplo, cuando $p = 0.0125$, el tamaño de la muestra preliminar (r_p) es 48 y la probabilidad de obtener un $W \leq \omega = 0.007$ es 0.4738842. Con $p = 0.02$, $r_p = 123$, sólo podemos estar 49.12 % seguros de que W será $\leq \omega = 0.007$. Cuando el número de unidades positivas aumenta en 10 (r_{m10},cuarta columna) o 20 (r_{m20}, sexta columna), la probabilidad $P(W \leq \omega = 0.007)$ aumenta.

Por ejemplo, cuando $p = 0.0125$, hay $r_{m20} = 68$ unidades en la muestra con $P(W \leq \omega = 0.007) = 0.9114708$; para $r_{m40} = 88$ unidades positivas en la muestra, la $P(W \leq \omega = 0.007) = 0.9962598$. Por lo tanto, los resultados del Cuadro 1 muestran que para garantizar una alta $P(W \leq \omega = 0.007)$, se requiere un tamaño de muestra (número de unidades positivas) mayor que el valor r_p calculado con la Ec. 5. Además, el Cuadro 1 muestra que en los 9 casos el tamaño de la muestra preliminar (número de observaciones positivas) resultante del uso de la Ec. 5 produce una $P(W \leq \omega) < 0.50$, es decir, 100% de las veces $P(W \leq \omega = 0.007)$ es menor que 50%.

Caso de estimación de la prevalencia vegetal bajo muestreo inverso

Como ejemplo de estimación para la utilización de las técnicas mencionadas se desarrolla el siguiente caso:

Ejemplo 1. Suponga que un investigador está interesado en estimar la proporción de plantas infectadas con virus en una empresa agrícola, cuya población es de N= 4 300 plantas, se decide usar muestreo inverso bajo muestreo aleatorio simple (MAS). Dado que la prevalencia de plantas infectadas es baja se establece detener el proceso de muestreo hasta que se encuentren r=5 plantas infectadas. Además, se lleva el registro del total de plantas extraídas y analizadas. Es decir, se extraerá sin remplazo una planta y se analizará si está infectada. Este proceso de extracción continuará hasta que se encuentren 5 plantas infectadas. El número total de plantas analizadas hasta encontrar las 5 infectadas fue de n=250. Los cálculos se realizarán con una precisión de 10% (d= 10/100) de la proporción preliminar (p), una confiabilidad $(1 - \alpha)$ de 95% (α= 95/100) y un nivel de aseguramiento (γ) de 99% (γ= 99/100). En el Cuadro 2 se muestran los cálculos correspondientes para la estimación de varios parámetros de interés.

Del Cuadro 2 se tiene que la proporción de plantas infectadas es de p= 0.02 con 95% de confiabilidad se estima que la proporción verdadera está entre 0.003158 y 0.036842, es decir, entre 0.31 y 3.68%. Se estima que el total verdadero es de 86 plantas infectadas, y está entre 13.57 y 158.42.

Example 1. Suppose a researcher is interested in estimating the proportion of virus infected plants in an agricultural company, with a population of N= 4 300 plants, it is decided to use low single inverse sampling random sampling (MAS). Since the prevalence of infected plants is set low to stop the sampling process until they are r= 5 infected plants. In addition, the registration of all plants is carried extracted and analyzed. It is, without replacement plant extract and analyze if you are infected. This extraction process will continue until 5 are infected plants. The total number of analyzed to find the five plants was infected n= 250. The calculations are made with an accuracy of 10% (d= 10/100) of the preliminary ratio (p), reliability $(1 - \alpha)$ of 95% (α= 95/100) and a level of assurance (γ) of 99% (γ= 99/100). In the Table 2 shows the calculations for estimating several parameters of interest are shown.

Table 2 is that the proportion of infected plants is $p = 0.02$ with 95% reliability is estimated that the real ratio is 0.036842 and 0.003158, i.e. between 0.31 to 3.68%. It is estimated that the actual total is 86 infected plants, and between 13.57 and 158.42 is. Finally you have to sample sizes under the traditional approach, and under the AIPE approach are 71 and 73 respectively.

Clearly manually do the calculations in Table 2 I is slow, laborious and with a high probability of being wrong. Therefore, the need to develop a free software that allows obtaining these parameters quickly, effectively and safely considered.

Software for example reverse-sampling

The software sampling was raised from two purposes: a) to determine the sample size for a study under the scheme of inverse sampling ratio and b) the estimation of the parameters resulting from applying this sampling scheme.

To calculate a sample size there is a contradiction in having to meet in the design phase, the value of a parameter that can only be estimated once extracted from the sample. To solve this problem you have two options, the first is to estimate the parameter from a pilot sample; the second is to obtain an acceptable value through the references have jobs or similar experiences already made. In the software for reverse sampling was chosen option 1. In addition, it is also required to specify the level of significance (α) which is usually less value than 0.2 (20%) and accuracy, which is defined as estrangement maximum time allowed between

Finalmente se tiene que los tamaños de muestra bajo el enfoque tradicional y bajo el enfoque AIPE son de 71 y 73 respectivamente.

the parameter and its estimate. The software allows the user to provide the required accuracy (directly) or obtained indirectly as a percentage of the estimated proportion preliminary.

Cuadro 2. Cálculos manuales de los parámetros de interés.
Table 2. Manual calculations of the parameters of interest.

Cálculos	Operaciones
Proporción de plantas	$\hat{p}=\frac{r}{n}=\frac{5}{250}=0.02$ y $\hat{q}=1\ \hat{p}=1-0.02=0.98$.
Varianza y Desviación estándar de la proporción	$V(\hat{p})=\left[\frac{N-n}{N}\right]\left[\frac{p^2q}{r}\right]$; donde: N= 4,300, n= 250, r= 5, $\hat{p}$= 0.02 y $\hat{q}$=0.98. Por lo tanto: $V(\hat{p})=\left[\frac{4{,}300-250}{4{,}300}\right]\left[\frac{(0.02)^2(0.98)}{5}\right]=(0.94186)(0.0000784)=0.000074$. Desviación estándar: $\sqrt{\hat{V}(\hat{p})}=0.008593$.
IC para la proporción.	$p_L=\hat{p}-Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})}$; $p_U=\hat{p}-Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})}$ donde $\hat{p}=0.02$, $\sqrt{\hat{V}(\hat{p})}=0.008593$ y $Z_{1-\alpha/2}=Z_{1-\alpha/2}=1.96$. Por lo tanto: $p_L=0.003158$ y $p_U=0.036842$.
Total	$\hat{\tau}=N\hat{p}$; donde: N= 4,300 y $\hat{p}$= 0.02. Por lo tanto $\hat{\tau}=(4{,}300)(0.02)=86$
IC para el total	$p_L=\hat{\tau}-N\,Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})}$; $p_U=\hat{\tau}+N\,Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{p})}$ donde $\hat{\tau}=86$, N= 4,300, $\sqrt{\hat{V}(\hat{p})}=0.008593$ y $Z_{1-\alpha/2}=1.96$. Por lo tanto: $p_L=13.578196$ y $p_U=158.421804$.
Tamaño de muestra para estimar la proporción	$r^*=\frac{N(Z_{1-\alpha/2})^2\hat{p}^2\hat{q}}{Nd^2+(Z_{1-\alpha/2})^2\hat{p}\hat{q}}$; donde: N= 4,300, $Z_{1-\alpha/2}$= 1.96, $\hat{p}$= 0.02 y $\hat{q}$=0.98 y d= 0.10(p)= 0.10(0.02)= 0.002. Por lo tanto: $r^*=\frac{(4{,}300)(1.96)^2(0.02)^2(0.98)}{(4{,}300)(0.002)^2+(1.96)^2(0.02)(0.98)}=\frac{6.475401}{0.092495}=70.0082=71$.
Tamaño de muestra modificado (AIPE) para estimar la proporción	Primeramente se obtendrá r_{MI} suponiendo una población infinita: $r_{MI}=\left[\frac{Z_{1-\alpha/2}}{\omega}\right]^2\left[\sqrt{\hat{p}^2\hat{q}}+\sqrt{p^2\,\hat{q}+\frac{2\omega\lvert(1.5\hat{p}^4-\hat{p}^3)\rvert Z_\gamma}{Z_{1-\alpha/2}\,\hat{p}^2}}\right]^2$; donde: N= 4,300, $Z_{1-\alpha/2}\,\hat{p}^2$ 1.96, $\hat{p}$= 0.02, $\hat{q}$= 0.98, Z_γ= 2.33 y ω= 2d= 2(0.002)= 0.004. Por lo tanto: $r_{MI}=\left[\frac{1.96}{0.004}\right]^2\left[\sqrt{(0.02)^2(0.98)}+\sqrt{(0.02)^2(0.98)+\frac{2(0.004)\lvert1.5(0.02)^4-(0.02)^3\rvert(2.33)}{1.96(0.02)^2}}\right]^2$ $r_{MI}=240{,}100[0.019799+0.02401]^2=460.806758$. $r_{MF}=r_{MI}\left[\frac{N}{N+r_{MI}/p}\right]$; donde: N= 4,300, r_{MI}= 460.806758 y $\hat{p}$= 0.02. Por lo tanto: $r_{MF}=460.806758\left[\frac{4{,}300}{4{,}300+460.806758/0.02}\right]=72.4742=73$.

Es evidente que hacer manualmente los cálculos del Cuadro 2 resulta tardado, laborioso y con una alta probabilidad de equivocarse. Por lo tanto, se consideró la necesidad de

However, since the sample sizes are calculated under the AIPE approach must also specify the level of assurance (γ) and this should be a value greater than 0.5 (50%).

elaborar un software de distribución libre que permita la obtención de estos parámetros de manera rápida, efectiva y segura.

Software-ejemplo para muestreo inverso

El software para muestreo se planteó a partir de dos propósitos: a) determinar el tamaño de muestra para un estudio bajo el esquema de muestreo inverso para una proporción y b) realizar la estimación de los parámetros resultantes de aplicar este esquema de muestreo.

Para el cálculo de un tamaño de muestra hay una contradicción al tener que conocer, en la fase de diseño, el valor de un parámetro que sólo se podrá estimar una vez extraída la muestra. Para resolver este problema se cuenta con dos opciones, la primera es estimar el parámetro a partir de una muestra piloto; la segunda es obtener un valor aceptable a través de las referencias que se tengan de trabajos o experiencias similares ya realizados. En el software para muestreo inverso se optó por la opción 1. Además, también se requiere especificar el nivel de significancia (α) el cual normalmente es un valor menor a 0.2 (20%) y la precisión, la cual se define como el alejamiento máximo permitido entre el parámetro y su estimación. El software permite al usuario brindar su precisión requerida (forma directa) u obtenerla en forma indirecta como un porcentaje con respecto a la proporción preliminar estimada. Sin embargo, dado que los tamaños de muestra se calculan bajo el enfoque AIPE también se debe de especificar el nivel de aseguramiento (γ) y este debe ser un valor mayor a 0.5 (50%).

A continuación se ilustra la utilización del software resolviendo el Ejemplo 1. Primeramente se introducen los datos solicitados por el software y se da clic en calcular como se ilustra en la Figura 1:

No hay que perder de vista que la muestra de 250 es una muestra piloto que sólo sirve para obtener la muestra definitiva. En este caso se puede observar en el área de resultados (Figura 1) que el tamaño de muestra modificado (definitivo), es de 73 plantas. Este tamaño de muestra garantiza que se cumplirá la precisión con una certidumbre de 99%.

Es importante resaltar que este tamaño de muestra es mayor al tamaño de muestra estimado de manera tradicional que es 71 (Figura 1), el cual sólo asegura que se cumpla la precisión requerida en un 50% de las veces, debido a que no toma en cuenta que el parámetro estimado con la muestra

Then using the software is illustrated by solving Example 1. First the data requested by the software are introduced and click on calculate as illustrated in Figure 1.

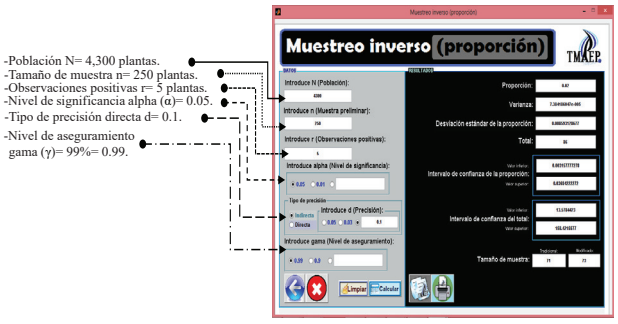

Figura 1. Interfaz del muestreo inverso para proporción.
Figure 1. Interface inverse proportion sampling.

We should not lose sight that the sample of 250 is a pilot shows that only serves to obtain the final sample. In this case can be seen in the results area (Figure 1) modified size sample (final), it is 73 plants. This sample size ensures that the accuracy will be fulfilled with a certainty of 99%.

Importantly, this sample size is greater than the estimated sample size traditionally is 71 (Figure 1), which only ensures that the required accuracy, 50% of the time is met, because does not take into that the preliminary estimated sample parameter is an estimate. Therefore, it is clear that the sample size under the AIPE approach (modified sample size) is usually greater to ensure the specified level of assurance.

Finally the researcher will estimates of the parameters of interest (percentages, totals, IC's) with the final sample using the same software but using the information collected from the final sample. That is, the preliminary sample is used only to obtain the final sample and to verify possible problems in the implementation of the survey. Therefore, the researcher must conduct the survey again but stop the sampling process until you find 73 (size modified sample) positive elements and the information gathered can use the same software for estimating parameters of interest which shall be valid estimates for the entire study population.

It is noteworthy that the software is not limited to the inverse sampling, but can also be used for sampling: simple, systematic, random stratified random, cluster at one stage, to estimate an average or proportion and even with methods such

preliminar es una estimación. Por lo tanto, es claro que el tamaño de muestra bajo el enfoque AIPE (tamaño de muestra modificado) es normalmente mayor para garantizar el nivel de aseguramiento especificado.

Finalmente el investigador hará las estimaciones de los parámetros de su interés (porcentajes, totales, IC's) con la muestra definitiva utilizando el mismo software pero usando la información recabada de la muestra definitiva. Es decir, la muestra preliminar sólo se utiliza para obtener la muestra definitiva y para verificar posibles problemas en la aplicación de la encuesta. Por lo tanto, el investigador debe realizar la encuesta nuevamente pero parar el proceso de muestreo hasta que encuentre a 73 (tamaño de muestra modificado) elementos positivos y con esta información recabada puede utilizar el mismo software para realizar la estimación de los parámetros de su interés que serán estimaciones válidas para toda la población bajo estudio.

Es importante mencionar que el software no está limitado al muestreo inverso, sino que también se puede usar para los muestreos: aleatorio simple, sistemático, aleatorio estratificado, por conglomerados en una etapa, para estimar un promedio o proporción y aún con métodos como respuesta aleatorizada (versión Horvitz) y pruebas de grupo (Group testing). Se puede consultar el manual detallado de todas las capacidades del software en https: //dl.dropboxusercontent.com/u/97440566/Manual%20TMAPEP.pdf.

Conclusiones

La fórmula presentada para el cálculo del tamaño de muestra para estimar la prevalencia de plantas bajo muestreo inverso con enfoque AIPE garantiza que con una probabilidad γ (nivel de aseguramiento ≥ 0.5) la precisión fijada a priori del IC se cumpla. Lo cual produce mayor exactitud en el estudio de interés realizado.

El software diseñado de distribución libre para muestreo inverso es una excelente herramienta que permite a investigadores y estudiantes lograr estimaciones exactas en los parámetros de su interés y se recomienda usar cuando la proporción a estimar es rara. Si bien la presentación de los casos mostrados es limitada, puesto que, sólo se proporciona as randomized response (Horvitz version) and test group (group testing). You can consult the detailed manual log of all software capabilities in https: //dl.dropboxusercontent.com/u/97440566/manual%20tmapep.pdf.

Conclusions

The formula presented for the calculation of sample size to estimate the prevalence of plants under reverse AIPE sampling approach ensures that with a probability γ (assurance level ≥ 0.5) fixed a priori accuracy of the IC is met. Which produces more accurate study performed interest.

The free distribution of software designed to reverse sampling is an excellent tool that enables researchers and students to achieve accurate estimates on the parameters of interest and is recommended when the ratio estimate is rare. While presenting the cases shown is limited, since only an example for the scheme inverse sampling is provided, this software can be used for on time or interval estimate an average, a proportion or a total under simple random sampling, stratified, systematic cluster and a stage and even for highly specialized sampling schemes such as reverse sampling and testing randomized response group. Although only one type of use of the software is illustrated clearly is used both for determining the sample size and for estimating parameters of interest using a pilot sample. Although the software requires the user to consider a pilot sample, and determine the final sample size, this ensures quality in the final estimates.

For the researcher or student succeeds in using the software fully, not only to reverse sampling should understand what this sampling scheme and its characteristics, as well as having clear parameters to estimate for the corresponding study. Therefore, if all decisions at this stage are justified properly, the software developed will be of great help, because, from the point of view statistical and computational software is very reliable.

End of the English version

un ejemplo para el esquema de muestreo inverso, este software puede ser utilizado para estimar puntualmente o por intervalo un promedio, una proporción o un total bajo muestreo aleatorio simple, estratificado, sistemático y por conglomerados en una etapa y aún para esquemas de muestreo muy especializados como son muestreo inverso, respuesta aleatorizada y pruebas de grupo. A pesar de que sólo se ilustra un tipo de utilización del software es evidente que se utiliza tanto para la determinación del tamaño de muestra, así como para la estimación de los parámetros de interés utilizando una muestra piloto. Aunque el software obliga al usuario considerar una muestra piloto, y así determinar el tamaño de muestra definitivo, esto garantiza calidad en las estimaciones definitivas.

Para que el investigador o estudiante tenga éxito al usar el software cabalmente, no sólo para muestreo inverso debe entender en qué consiste este esquema de muestreo y sus características, así como también tener claros los parámetros que desea estimar para su correspondiente estudio. Por ello, si todas las decisiones en esta fase son justificadas de manera adecuada, el software desarrollado le será de gran ayuda, debido a que, desde el punto de vista estadístico y computacional el software es muy confiable.

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. 45:969-977.

Czaja, R.; Snowden, C. B. and Casady, R. 1996. Reporting vias and sampling errors in a survey of a rare population using multiplicity counting rules. Journal of the American Statistical Association. 81(394):411-419.

Fosgate, G. T. 2007. A cluster-adjusted sample size algorithm for proportions was developed using a beta-binomial model. Journal of clinical epidemiology. 60(3):250-255.

George, V. T. and Elston, R. C. 1993. Confidence limits based on the first occurrence of an event. Statistics in Medicine. 12(7):685-690.

Graham, K. and Dallas, A. 1988. Sampling rare populations. J. Royal Statistical Soc. 149:65-82.

Haldane, J. B. 1945. On a method of estimating frequencies. Biometrika. 33(3):222-225.

Hernández-Suárez, C. M.; Montesinos-López, O. A.; McLaren, G. and Crossa J. 2008. Probability models for detecting transgenic plants. Seed Sci. Res. 18(02):77-89.

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

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

Kelly, K. and Rausch, J. R. 2011. Sample size planning for longitudinal models: Accuracy in parameter estimation for polynomial change parameters. Psychological Methods. 16(4):391-405.

Krebs, C. J. 2001. Ecology. The experimental analysis of distributions and abundance. Harper and Row publishers. Michigan, USA. 801 p.

Lui, K. J. 2001. Estimation of rate ratio and relative difference in matched-pairs under inverse sampling. Environmetrics. 12(6):539-546.

Madden, L. V.; Hughes, G. and Munkvold, G. P. 1996. Plant disease incidence: inverse sampling, sequential sampling, and confidence intervals when observed mean incidence is zero. Elsevier. 15(7):621-632.

Montesinos, L. O. A.; Montesinos, L. A.; Crossa, J. and Kent, E. 2012. Sample size under inverse negative binomial group testing for accuracy in parameter estimation. Plos One. 7(3):1-10.

Montesinos, L. O. A.; Montesinos, L. A.; Crossa, J.; Eskridge, K. and Sáenz, C. 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.

Oehlert, G. W. 1992. A note on the delta method. The American Statistician. 46(1):27-29.

Ragan, V. E. 2002. The animal and plant health inspection service (APHIS) brucellosis eradication program in the United States. Veterinary Microbiol. 90(1):11-18.

Sheaffer, R. L. and Leavenworth, R. S. 1976. The negative binomial model for counts in units of varying size. J. Quality Technol. 8(3):158-163.

Singh, P. and Aggarwal, A. R. 1991. Inverse sampling in case control studies. Environmetrics. 2(3):293-299.

Sudman, S.; Sirken, M. and Cowan C. 1988. Sampling rare and elusive populations. Science. 240(4855):991-996.

Tang, M. L.; Liao, Y. J. and Ng, H. K. T. 2008. On test of rate ratio under standard inverse sampling. Computer methods and programs in biomedicine 89(3):261-268.

Zhu, H. and Lakkis, H. 2014. Sample size calculation for comparing two negative binomial rates. Statistics in Medicine. 33(3):376-387.