Agrociencia
ISSN: 1405-3195
agrocien@colpos.mx
Colegio de Postgraduados
México

Reyes Carreto, Ramón; Ramírez Valverde, Gustavo

Prueba Bootstrap para la hipótesis de no preferencia en estudios toxicológicos con variables dicotómicas

Agrociencia, vol. 36, núm. 3, mayo-junio, 2002, pp. 329-335
Colegio de Postgraduados
Texcoco, México

PRUEBA BOOTSTRAP PARA LA HIPÓTESIS DE NO PREFERENCIA EN ESTUDIOS TOXICOLÓGICOS CON VARIABLES DICOTÓMICAS

BOOTSTRAP TEST FOR THE NON-PREFERENCE HYPHOTESIS IN EXPERIMENTS WITH DICHOTOMOUS VARIABLES

Ramón **Reyes-Carreto**[1] y Gustavo **Ramírez-Valverde**[2]

[1]Facultad de Matemáticas. Universidad Autónoma de Guerrero. Ciudad Universitaria. 39000. Chilpancingo, Guerrero. [2]Especialidad de Estadística. Instituto de Socioeconomía, Estadística e Informática. Colegio de Postgraduados. 56230. Montecillo, Estado de México. (gramirez@colpos.mx)

Resumen

En estudios toxicológicos, con frecuencia se prueban la preferencia o no-preferencia de insectos en experimentos con variable de respuesta dicotómica. Este tipo de experimento, cuando se supone la presencia de sobre-dispersión, se ha modelado usando la distribución Beta-binomial. Se puede construir la prueba de razón de verosimilitud para probar la no-preferencia $\left(\mathbf{H_0}{:}\pi = \frac{1}{2}\right)$; sin embargo, para muestras pequeñas, no se logra una buena aproximación a la distribución asintótica del estadístico. Este trabajo tuvo como objetivo diseñar una técnica no paramétrica de Bootstrap sobre el estadístico de razón de verosimilitud, en un modelo Beta-Binomial, para probar la no preferencia en experimentos con dos opciones. La prueba de hipótesis propuesta utiliza como estadístico de prueba el de razón de verosimilitud $\left(\hat{\delta}\right)$. Mediante la técnica Bootstrap se estima la distribución muestral $\left(\hat{F}\left(\hat{\delta}\right)\right)$ bajo $\mathbf{H_0}$. Finalmente, para probar la hipótesis de no preferencia $\left(\mathbf{H_0}{:}\pi = \frac{1}{2}\right)$ con un nivel de significancia (α) determinado, se rechaza $\mathbf{H_0}$ si $\hat{\delta} = -2\ln\hat{\lambda} > q_{(1-\alpha)}$, donde $q_{(1-\alpha)}$ es el percentil Bootstrap $(1-\alpha)$x100. Para ilustrar esta prueba se elaboró un programa con IML de SAS, y mediante un ejemplo se muestra el uso de la metodología propuesta. Se concluye que la técnica Bootstrap permite obtener una aproximación a la distribución del estadístico de razón de verosimilitud usando el modelo Beta-Binomial, cuando el experimento involucra variables dicotómicas y los datos son afectados por la sobredispersión.

Palabras claves: Beta-Binomial, Bootstrap, distribución muestral, método Newton-Raphson, sobredispersión.

Introducción

Frecuentemente en estudios toxicológicos se plantea una prueba de hipótesis sobre la preferencia o no preferencia de insectos en experimentos con

Recibido: Noviembre, 1999. Aprobado: Febrero, 2002.
Publicado como ENSAYO en Agrociencia 36: 329-335. 2002.

Abstract

In toxicological studies, the preference or non-preference of insects in experiments with a dichotomous response variable is frequently tested. This type of experiment, when over-dispersion is assumed, has been modeled using the Beta-Binomial distribution. Hypotheses of non-preference $\left(\mathbf{H_0}{:}\pi = \frac{1}{2}\right)$; can be tested using the generalized likelihood ratio; however, for small samples, a good approximation to the asymptotic distribution of the statistic is not achieved. This research had the objective of designing a non-parametric Bootstrap technique on the likelihood ratio statistic in a Beta-Binomial model, to test non-preference in experiments with two choices. The proposed test of hypothesis uses as test criterion the likelihood ratio $\left(\hat{\delta}\right)$. Using the Bootstrap technique, the sample distribution $\left(\hat{F}\left(\hat{\delta}\right)\right)$ is estimated under $\mathbf{H_0}$. Finally, to test the hypothesis of non-preference $\left(\mathbf{H_0}{:}\pi = \frac{1}{2}\right)$, with a given level of significance (α), $\mathbf{H_0}$ is rejected if $\hat{\delta} = -2\ln\hat{\lambda} > q_{(1-\alpha)}$, were $q_{(1-\alpha)}$ is the $(1-\alpha)$x100 Bootstrap percentile. To illustrate this test, a program with SAS' IML was created, and the proposed methodology is exemplified. It is concluded that the Bootstrap technique permits an approximation to the likelihood ratio statistic distribution, using the Beta-Binomial model, when the experiment involves dichotomous variables and data are affected by over-dispersion.

Key words: Beta-binomial, Bootstrap, sampling distribution, Newton-Raphson method, over-dispersion.

Introduction

Frequently, in toxicological studies, the aim is to test a hypothesis on the preference or non-preference of insects in experiments with dichotomous response variable. For example, places to deposit eggs, election of mates, or types of food; or else a certain number of insects are placed in conditions to freely select one of two possible types of fruit for ovipositing variable de respuesta dicotómica. Por ejemplo, lugares para depositar huevecillos, elección de parejas, o tipos de alimentos; o bien, a cierto número de insectos se les pone en condiciones de seleccionar libremente uno de dos posibles tipos de frutos para ovipositar. Vaillant y Derridj (1992). El resultado del experimento se clasifica como éxito o fracaso. Si el experimento se realiza *n* veces de manera independiente, y la probabilidad de éxito en cada repetición es *p*, entonces la variable aleatoria número total de éxitos en las *n* realizaciones tendrá distribución Binomial con parámetros *n* y *p*. Se quiere estimar la proporción real de individuos, *p*, que resultaron éxitos.

El presente trabajo tuvo como objetivo diseñar una técnica no paramétrica Bootstrap para este problema en presencia de sobredispersión. Particularmente, se propone estimar la distribución exacta de la razón de verosimilitud usando Bootstrap, en vez usar la distribución asintótica.

El problema de sobredispersión

Al repetir *m* veces el experimento binomial, generalmente existen factores biológicos o cambios en las condiciones ambientales que alteran el comportamiento de los insectos, variando el parámetro *p* en cada repetición. Cuando esto sucede los supuestos originales no se cumplen. La consecuencia de que varíe la probabilidad de éxito, *p*, en cada realización del experimento, provoca el fenómeno de sobredispersión. Así, este problema se presenta cuando los datos muestran mayor variabilidad que la predicha por el modelo binomial supuesto.

En general, la sobredispersión puede surgir por varias razones. En este trabajo interesa la causada por probabilidades aleatorias, que se presenta cuando se tienen grupos de ensayos binomiales y la probabilidad de éxito varía entre los grupos, pese a estar en condiciones similares.

Efecto de las probabilidades en las varianzas

Sea Y_i una variable binaria observada en *n* individuos. Si las Y_i son independientes y $Y=\sum_{i=1}^{n} Y_i$ es el número total de éxitos, *Y* es una variable aleatoria con distribución Binomial con parámetros *n* y *p*. Si el valor del parámetro binomial, *p*, es una realización de la variable aleatoria Π con función de densidad $f_{\Pi}(p)$, $(0 \leq p \leq 1)$, la esperanza y varianza condicionales de *Y* son, $E(Y/\Pi=p)=np$ y $\mathrm{Var}(Y/\Pi=p)=np(1-p)$; mientras que la media y varianza no condicionales de *Y* son $E(Y)=E\left[E(Y/\Pi=p)\right]=E(n\Pi)=n\mu_{\pi}$ y

(Vaillant and Derridj, 1992). The result of the experiment is classified as success or failure. If the experiment is carried out independently *n* times and the probability of success in each replication is *p*, then the random variable total number of successes in the *n* trials will have a Binomial distribution with parameters *n* and *p.* The goal is to estimate the real proportion of individuals, *p*, that were successes.

This work had the objective of designing a non-parametric Bootstrap technique for this problem in presence of over-dispersion. Particularly, it is proposed to estimate the exact distribution of the likelihood ratio using Bootstrap, instead of the asymptotic distribution.

The problem of over-dispersion

When the binomial experiment is repeated *m* times, there generally exist biological factors or changes in environmental conditions that alter the behavior of the insects, and parameter *p* varies in each replication. When this occurs, the original assumptions are not valid. The consequence of the probability of success, *p*, varying every time the experiment is carried out causes the phenomenon of over-dispersion. Thus, this problem is present when the data show greater variability than the one predicted by the assumed Binomial model.

In general, over-dispersion can occur for several reasons. This study focuses on that caused by random probabilities, which occurs when groups of binomial trials are involved and the probability of success varies among groups, in spite of their similar conditions.

Effect of the probabilities in the variances

Let Y_i be a dichotomous variable observed in *n* individuals. If the Y_i are independent, and $Y=\sum_{i=1}^{n} Y_i$ is the total number of successes, then *Y* is a random variable with a Binomial distribution with parameters *n* and *p*. If the value of the binomial parameter, *p*, is a realization of the random variable Π with a density function $f_{\Pi}(p)$, $(0 \leq p \leq 1)$, the conditional expectation and variance of *Y* are $E(Y/\Pi=p)=np$ and $\mathrm{Var}(Y/\Pi=p)=np(1-p)$; while the non-conditional mean and variance of *Y* are and $E(Y)=E\left[E(Y/\Pi=p)\right]=E(n\Pi)=n\mu_{\pi}$, where μ_{π} and σ_{π}^2 are the mean and variance of the Π distribution.

If the variance of Π is defined by $\sigma_{\pi}^2=\phi\mu_{\pi}(1-\mu_{\pi})$ with $\phi>0$, $\mathrm{Var}(Y)=n\mu_{\pi}(1-\mu_{\pi})\{1+(n-1)\phi\}$. Factor $\{1+(n-1)\phi\}>1$ is called the over-dispersion factor.

$\mathrm{Var}(Y) = n\mu_\pi(1-\mu_\pi) + n(n-1)\sigma_\pi^2$; donde μ_π y σ_π^2 son la media y varianza de la distribución Π.

Si la varianza de Π se define por $\sigma_\pi^2 = \phi\mu_\pi(1-\mu_\pi)$ con $\phi > 0$, entonces $\mathrm{Var}(Y) = n\mu_\pi(1-\mu_\pi)\{1+(n-1)\phi\}$. Al factor $\{1+(n-1)\phi\} > 1$ se le llama factor de sobre-dispersión.

Para el análisis de este tipo de experimentos se han aplicado técnicas estadísticas como la *t* para muestras apareadas, el análisis de varianza y la prueba de Wilcoxon con signos. La aplicación de las pruebas de *t* y de Wilcoxon presenta el inconveniente de que si existe efecto de bloque, habrá interacción bloques por tratamientos, interacción que pone en duda la aplicación adecuada de estas técnicas. Méndez y Ramírez (2001), utilizan el modelo Beta-Binomial y proponen la prueba de razón de verosimilitud, concluyendo que dicho modelo proporciona mejores estimaciones del tamaño de prueba (α=0.05) y mayores potencias respecto a otras pruebas. Sin embargo, observaron algunos casos donde no se tuvo buena aproximación al tamaño de la prueba supuesto. Esto se debe a que la prueba de razón de verosimilitud es asintótica.

El Método Beta-Binomial

Sean $y_1, y_2,...,y_m$ variables aleatorias independientes que representan el número de éxitos en *m* experimentos binomiales con tamaño *n* cada uno. Si p_i es la probabilidad de éxito en cada experimento y suponemos que p_i tiene distribución Beta, entonces, $f(y/p) = \binom{n}{y} p^y(1-p)^{n-y}$, y=0,1,...*n*. Y puesto que la distribución Beta es $g(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$, 0<*p*<1, *a*>0, *b*>0, entonces la distribución no condicional del número observado de éxitos *Y*, está dada por la expresión

$$f(y) = \int_0^1 f(y,p)dp = \int_0^1 f(y/p)g(p)$$

$$dp = \binom{n}{y} \frac{B(y+a, n+b-y)}{B(a,b)}, \; y=0,1,2,...n,\; a>0,\; b>0.$$

Esta función se conoce como distribución Beta-Binomial.

Usando la relación entre las funciones gamma y beta y con la reparametrización $\pi = \frac{a}{a+b}$ y $\theta = \frac{1}{a+b}$, la distribución de *Y* puede reescribirse como $P[Y=y] = \binom{n}{y} \frac{\prod_{r=0}^{y-1}(\pi + r\theta) \prod_{r=0}^{n-y-1}(1-\pi+r\theta)}{\prod_{r=0}^{n-1}(1+r\theta)}$ donde,

To analyze this type of experiments statistical techniques such as the paired *t* and the Wilcoxon signed-rank test have been applied. The application of *t* and the Wilcoxon test has the disadvantage that, if there is a block effect, there will be an interaction between blocks and treatments, an interaction that places doubt on the appropriate application of these techniques. Méndez and Ramírez (2001) used the Beta-Binomial model and proposed the likelihood ratio test. They conclude that this model provides better estimations of the test size (α=0.05) and is more powerful respect to other tests. However, they found some cases where a good estimation of the assumed test size was not achieved. This is due to the fact that the likelihood ratio test is asymptotic.

The Beta-Binomial Method

Let be independent random variables that represent the number of successes in *m* binomial experiments, each of size *n*. If p_i is the probability of success in each experiment, and we assume that p_i has a Beta distribution, then $f(y/p) = \binom{n}{y} p^y(1-p)^{n-y}$, y=0,1,...*n*. And, since the Beta distribution is $g(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$, 0<*p*<1, *a*>0, *b*>0, then the non-conditional distribution of the observed number of successes *Y* is given by the expression

$$f(y) = \int_0^1 f(y,p)dp = \int_0^1 f(y/p)g(p)$$

$$dp = \binom{n}{y} \frac{B(y+a, n+b-y)}{B(a,b)}, \; y=0,1,2,...n,\; a>0,\; b>0.$$

This function is known as the Beta-Binomial distribution.

Using the relationship between gamma and beta functions, and with the reparameterization $\pi = \frac{a}{a+b}$ and $\theta = \frac{1}{a+b}$, the distribution of *Y* can be rewritten as $P[Y=y] = \binom{n}{y} \frac{\prod_{r=0}^{y-1}(\pi + r\theta) \prod_{r=0}^{n-y-1}(1-\pi+r\theta)}{\prod_{r=0}^{n-1}(1+r\theta)}$ where 0<π<1, y $\theta \geq 0$. The non-conditional mean and variance of the random variable *Y* are given by $E(Y) = n\pi$ and $\mathrm{Var}(Y) = \frac{n\pi(1-\pi)(1+n\theta)}{1+\theta}$.

With this reparameterization π is the expected value of *p*, and θ represents the variation of *p*. Also, if θ=0, the model is reduced to a Binomial model.

$0<\pi<1$, y $\theta \geq 0$. La media y varianza no condicionales de la variable aleatoria Y están dadas por $E(Y)=n\pi$ y $\mathrm{Var}(Y)=\frac{n\pi(1-\pi)(1+n\theta)}{1+\theta}$.

Con esta reparametrización π es el valor esperado y θ la variación de p. También, si θ=0, el modelo se reduce a un modelo Binomial.

Prueba de razón de verosimilitud

Considérese una muestra aleatoria de la distribución Beta-Binomial con parámetros π y θ. Se desea probar la no preferencia, esto es, H_0: $\pi=\frac{1}{2}$ *versus* H_a: $\pi \neq \frac{1}{2}$. Para obtener los estimadores $\tilde{\theta}$ y $(\hat{\pi},\hat{\theta})$ se utiliza el método de máxima verosimilitud sobre el conjunto $\Omega_0=\left\{(\pi,\theta)\middle|\pi=\frac{1}{2},\theta\geq 0\right\}$ bajo H_0, y sobre $\Omega=\{(\pi,\theta)|0<\pi<1,\theta\geq 0\}$ bajo H_a. Debido a que el sistema de ecuaciones simultáneas no tiene una solución analítica, los estimadores se obtienen mediante el método Newton-Raphson.

Sea $\hat{\lambda}=\frac{L(\tilde{\theta})}{L(\hat{\prod},\hat{\theta})}$ el cociente de verosimilitud, donde $\tilde{\theta}$ es el estimador de máxima verosimilitud bajo H_0 y $\hat{\Pi}$, $\hat{\theta}$ son los estimadores de máxima verosimilitud en el espacio paramétrico irrestricto. Puede demostrarse que asintóticamente, bajo condiciones en las cuales los estimadores de máxima verosimilitud resultan ser los mejores, el estadístico de prueba $\hat{\delta}=-2\log\hat{\lambda}$, tiene una distribución Ji-cuadrada con un grado de libertad. Así, se rechaza H_0, si y sólo si, $\hat{\delta}\geq\chi^2_{1-\alpha}(1)$, con nivel de significancia α, donde $\chi^2_{1-\alpha}(1)$ es el cuantil $(1-\alpha)$ de la distribución Ji-cuadrada con 1 grado de libertad.

El modelo Beta-Binomial toma en cuenta la variabilidad debida a la sobredispersión; sin embargo, al obtener las estimaciones de los parámetros, el método Newton-Raphson puede no converger. Ello sucede sobre todo en ausencia de sobredispersión (Bailey, 1957).

Metodología Bootstrap

La técnica Bootstrap, originalmente propuesta por Efron (1979) es una técnica estadístico-computacional no-paramétrica para obtener inferencias sobre el parámetro θ mediante el estadístico $\hat{\theta}$, y no requiere supuestos distribucionales. La técnica Bootstrap puede aplicarse en muchas situaciones, particularmente cuando no se conoce o no puede suponerse la distribución muestral del estadístico o en aquellos casos en que la distribución muestral no tiene solución analítica o es intratable.

La justificación teórica está basada en dos consideraciones: 1) La Función de Distribución Empírica, $\hat{F}(y)$, estima a la Función de Distribución verdadera, $F(y)$; y el teorema Glivenko-Cantelli muestra que $\hat{F}(y)$ converge en probabilidad a $F(y)$. (Bickel y Freedman, 1981). Intuitivamente, cuando se incrementa el tamaño de la muestra, ésta contiene mayor información acerca de la población, y para *n=N*, $\hat{F}(y)=F(y)$. 2) La propiedad de consistencia permite a la distribución muestral Bootstraping $\hat{F}*(\hat{\theta}*)$ aproximar a $F(\hat{\theta})$ de una muestra dada, cuando el número de remuestreos B es grande y permite aproximar $\hat{F}(y)$ a $F(y)$. Bajo tales condiciones, y cuando el número

Likelihood ratio test

Consider a random sample of the Beta-Binomial distribution with parameters π and θ. Non-preference is to be tested, that is H_0: $\pi=\frac{1}{2}$ *vs* H_a: $\pi\neq\frac{1}{2}$. To obtain the estimators $\tilde{\theta}$ and $(\hat{\pi},\hat{\theta})$, the maximum likelihood method is used on the set $\Omega_0=\left\{(\pi,\theta)\middle|\pi=\frac{1}{2},\theta\geq 0\right\}$ under H_0, and on $\Omega=\{(\pi,\theta)|0<\pi<1,\theta\geq 0\}$ under H_a. Since the system of simultaneous equations does not have an analytical solution, the estimators are obtained by means of the Newton-Raphson method.

Let $\hat{\lambda}=\frac{L(\tilde{\theta})}{L(\hat{\prod},\hat{\theta})}$ be the likelihood ratio, where $\tilde{\theta}$ is the maximum likelihood estimator under H_0, and $\hat{\Pi}$, $\hat{\theta}$, are the maximum likelihood estimators in the unrestricted parametric space. It can be demonstrated that asymptotically, under conditions in which the maximum likelihood estimators turn out to be the best estimators, the test statistic $\hat{\delta}=-2\log\hat{\lambda}$, has a Chi-square distribution with one degree of freedom. Thus, H_0 is rejected if, and only if, $\hat{\delta}\geq\chi^2_{1-\alpha}(1)$, with an α level of significance, where $\chi^2_{1-\alpha}(1)$ is the quantile $(1-\alpha)$ of the Chi-square distribution with one degree of freedom.

The Beta-Binomial model takes into account the variability due to over-dispersion. However, when estimations of the parameters are obtained, the Newton-Raphson may not converge. This occurs especially in absence of over-dispersion (Bailey, 1957).

The Bootstrap Methodology

The Bootstrap technique, originally proposed by Efron (1979) is a non-parametric statistical-computational technique used to obtain inferences on the θ parameter using the $\hat{\theta}$ statistic, and does not require distributional assumptions. The Bootstrap technique can be applied to diverse situations, especially when the sample distribution of the statistic is unknown or cannot be assumed, or in those cases in which the sample distribution does not have an analytical solution or is untreatable.

The theoretical justification is based on two considerations: 1) The Empirical Distribution Function, $\hat{F}(y)$, estimates the true Distribution Function, $F(y)$; and the Glivenko-Cantelli theorem shows that $\hat{F}(y)$ converges in probability to $F(y)$ (Bickel and Freedman, 1981). Intuitively, when the sample size increases, it contains more information about the population, and for *n=N*, $\hat{F}(y)=F(y)$. 2) The consistency property permits a Bootstrapping sample distribution $\hat{F}*(\hat{\theta}*)$ to approximate $F(\hat{\theta})$ of a given sample, when the number of re-samplings B is large and allows to approximate $\hat{F}(y)$ to $F(y)$. Under such conditions, and when the number of re-samplings, *B*, is large enough, Babu and Singh (1983) demonstrate that $\hat{F}*(\hat{\theta}*)\approx F(\hat{\theta})$.

The Bootstrapping Procedure

1) The independent and identically distributed random variables Y_i's, *i*=1,2,..., *n* are defined as the random sample, with a distribution

de remuestreos, B, es suficientemente grande, Babu y Singh (1983) demuestran que $\hat{F}*(\hat{\theta}*) \approx F(\hat{\theta})$.

El procedimiento Bootstrap

1) Se define como muestra aleatoria a las variables aleatorias Y_i; i=1,2,..., n independientes e idénticamente distribuidas, con función de distribución $F(\cdot)$, que en este caso será la Beta-Binomial; 2) Se obtiene un remuestreo o una muestra Bootstrap $Y^* = \{y_1^*,...,y_n^*\}$, muestreando aleatoriamente n veces, con reemplazo, los datos originales y_1,...,y_n; el tamaño de la muestra aleatoria es el mismo de la muestra Bootstrap, y las Y_i^* tienen probabilidad n^{-1}, siendo igual a cada una de las Y_i; 3) Se calcula el estadístico de interés $\hat{\theta}$ de este remuestreo, produciendo $\hat{\theta}^*$ (para este caso será Ji-cuadrada); 4) Se repiten los pasos 1 y 2, B veces, para este ejemplo se usó B=1000; y 5) Se construye la distribución de probabilidad de las B $\hat{\theta}^*$, asignando probabilidad 1/B a cada $\hat{\theta}_i^*$. Esta es la estimación de la distribución muestral de $\hat{\theta}$, $\hat{F}*(\hat{\theta}*)$.

El intervalo percentil Bootstrap

Si se considera que $\hat{F}*(\hat{\theta}*)$ estima a $\hat{F}(\hat{\theta})$, la justificación teórica es simple. Un intervalo de confianza de nivel α incluye todos los valores de $\hat{\theta}^*$ entre los percentiles $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$ de la distribución $\hat{F}^*(\hat{\theta}^*)$ (Efron, 1982). Esto es, los límites de un intervalo de confianza de 95% para $\hat{\theta}$ son los valores de $\hat{\theta}^*$ en los percentiles 2.5 y 97.5 de $\hat{F}^*(\hat{\theta}^*)$.

Sea $\hat{G}$ la función de distribución acumulada de $\hat{\theta}^*$, el intervalo $1-2\alpha$ percentil se define por los percentiles α y $1-\alpha$ de $\hat{G}$ como:

$$\left(\hat{\theta}_{\text{inf}}, \hat{\theta}_{\text{sup}}\right) = \left(\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)\right) \quad (1)$$

Dado que por definición, $\hat{G}^{-1}(\alpha) = \hat{\theta}^{*(\alpha)}$ es el α-ésimo percentil de la distribución Bootstrap, se puede reescribir el intervalo percentil (1) como:

$$\left(\hat{\theta}_{\text{inf}}, \hat{\theta}_{\text{sup}}\right) = \left(\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}\right) \quad (2)$$

Las expresiones (1) y (2) se refieren a la situación ideal Bootstrap en la cual el número de repeticiones es infinito. En la práctica se debe utilizar algún número finito B de repeticiones.

Así, primero se procede a generar B conjuntos de datos Bootstrap $x^{*1}, x^{*2}, \ldots, x^{*B}$, y luego se calculan las repeticiones Bootstrap,

$$\hat{\theta}^*(b) = S(x^{*b}),\ b=1,2,...,B \quad (3)$$

Sea $\hat{\theta}^{*(\alpha)}$ el α-cuantil empírico de los valores $\hat{\theta}^*(b)$; esto es, el Bα-ésimo valor en la lista ordenada de las B repeticiones de $\hat{\theta}^*$. Por ejemplo, si B=2000 y α=0.05, $\hat{\theta}_{2000}{}^{*(.05)}$ es el 100-ésimo valor ordenado de las repeticiones. Asimismo, sea $\hat{\theta}_B{}^{*(1-\alpha)}$ el $(1-\alpha)$-ésimo percentil empírico. El intervalo $(1-2\alpha)$ percentil aproximado es

function, $F(\cdot)$, which, in this case will be the Beta-Binomial. 2) A re-sampling, or Bootstrap sample $Y^* = \{y_1^*,...,y_n^*\}$, is obtained, by sampling randomly n times the original data y_1,...,y_n, with replacement; the random sample is the same size as the Bootstrap sample, and the Y_i^* have probability n^{-1}, equal to each of the Y_i. 3) The statistic of interest $\hat{\theta}$ of this re-sampling is calculated, producing $\hat{\theta}^*$ (for this case it will be chi-square). 4) Steps 1 and 2 are repeated B times; for this example B=1000 was used; and 5) The probability distribution of the B $\hat{\theta}^*$s is constructed, assigning probability 1/B to each $\hat{\theta}_i^*$. This is the estimation of the sampling distribution of $\hat{\theta}$, $\hat{F}*(\hat{\theta}*)$.

The Bootstrap percentile interval

If it is assumed that $\hat{F}*(\hat{\theta}*)$ estimates $\hat{F}(\hat{\theta})$, the theoretical justification is simple. An α level confidence interval includes all $\hat{\theta}^*$ values between percentiles $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ of the $\hat{F}^*(\hat{\theta}^*)$ distribution (Efron, 1982). That is, the limits of a 95% confidence interval for $\hat{\theta}$ are the $\hat{\theta}^*$ values in the 2.5 and 97.5 percentiles of $\hat{F}^*(\hat{\theta}^*)$.

Let $\hat{G}$ be the cumulative distribution function of $\hat{\theta}^*$; the interval $1-2\alpha$ percentile is defined by the percentiles α and $1-\alpha$ of $\hat{G}$ as:

$$\left(\hat{\theta}_{\text{inf}}, \hat{\theta}_{\text{sup}}\right) = \left(\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)\right) \quad (1)$$

Given that, by definition, $\hat{G}^{-1}(\alpha) = \hat{\theta}^{*(\alpha)}$ is the αth percentile of the Bootstrap distribution, the percentile interval (1) can be rewritten as

$$\left(\hat{\theta}_{\text{inf}}, \hat{\theta}_{\text{sup}}\right) = \left(\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}\right) \quad (2)$$

Equations (1) and (2) refer to the ideal Bootstrap situation, in which the number of repetitions is infinite. In practice, some finite number B of repetitions must be used.

Thus, first B sets of Bootstrap data, $x^{*1}, x^{*2}, \ldots, x^{*B}$, are generated, then Bootstrap repetitions are calculated,

$$\hat{\theta}^*(b) = S(x^{*b}),\ b=1,2,...,B \quad (3)$$

Let $\hat{\theta}^{*(\alpha)}$ be the empirical α-quantile of the values $\hat{\theta}^*(b)$; that is, the Bαth value in the ordered list of the B repetitions of $\hat{\theta}^*$. For example, if B=2000 and α=0.05, $\hat{\theta}_{2000}{}^{*(.05)}$ is the 100th ordered value of the repetitions. Likewise, let $\hat{\theta}_B{}^{*(1-\alpha)}$ be the $(1-\alpha)$th empirical percentile. The approximate percentile interval $(1-2\alpha)$ is

$$\left(\hat{\theta}_{\text{inf}}, \hat{\theta}_{\text{sup}}\right) \approx \left(\hat{\theta}_B{}^{*(\alpha)}, \hat{\theta}_B{}^{*(1-\alpha)}\right) \quad (4)$$

The Bootstrap test for non-preference

This Bootstrap methodology has as its objective to test the hypothesis of non-preference H_0: $\pi = \frac{1}{2}$, with a given significance

$$\left(\hat{\theta}_{\inf},\hat{\theta}_{\sup}\right)\approx\left(\hat{\theta}_B{}^{*(\alpha)},\hat{\theta}_B{}^{*(1-\alpha)}\right) \quad (4)$$

Prueba Bootstrap para no preferencia

Esta metodología Bootstrap tiene por objeto realizar la prueba de hipótesis de no preferencia H_0: $\pi=\frac{1}{2}$, con un nivel de significancia determinado. El procedimiento consiste en: 1) De los datos originales de la muestra se obtienen los estimadores $\hat{\theta}$ y $\hat{\pi}$ del modelo Beta-Binomial, mediante la aproximación de Newton-Raphson; donde $\pi=\alpha/(\alpha+\beta)$ y $\theta=1/(\alpha+\beta)$. 2) En forma similar al paso 1) se obtiene el estimador del parámetro θ bajo H_0, y $\hat{\theta}_{H_0}$; y con ambos resultados se calcula el estadístico de razón de verosimilitud $\hat{\delta}=-2\log\hat{\lambda}$. 3) Se realizan 1000 muestras Bootstrap de la distribución Beta-Binomial bajo H_0: $\pi=\frac{1}{2}$ y $\hat{\theta}_{H_0}$. Cada remuestreo deberá ser del mismo tamaño que el de la muestra inicial. 4) Se calcula el estadístico $\hat{\delta}$ para cada muestra Bootstrap, y con ellos se construye la función de distribución empírica de $\hat{\delta}$. 5) La prueba de hipótesis de no preferencia, H_0: $\pi=\frac{1}{2}$, con α=0.05, equivale a obtener el percentil 95.0, $q_{0.95}$ del intervalo de confianza Bootstrap percentil, y rechazar H_0 si $\hat{\delta}>q_{0.95}$.

Resultados y Discusión

Para ilustrar esta metodología se elaboró un programa con IML de SAS[3], y se trabajó un ejemplo con tamaño n=20 para cada binomial, y un total de m=10 binomiales, usando los datos 8,19,13,1,7,20,15,13,8,20 y un total de B=1000 muestras Bootstrap. Se aplicaron los métodos de χ^2 cuadrada y de Fitzmaurice para detectar sobre-dispersión, y de los resultados se concluyó que el fenómeno de sobredispersión está presente en los datos del ejemplo (Méndez y Ramírez, 2001).

Con estos resultados la aplicación de la metodología propuesta resulta bastante racional. El programa se fundamenta en la equivalencia entre la prueba y el intervalo percentil Bootstrap. Así, la obtención del percentil 95.0 de la distribución empírica del estadístico $\hat{\delta}$, equivale a rechazar H_0 si $\hat{\delta}>q_{0.95}$. El estadístico y el valor critico al 5% de la prueba de razón de verosimilitud son $\hat{\delta}$ =1.8093 y 3.8414, respectivamente, por lo que no se rechaza la hipótesis nula de no preferencia de los insectos. Por otra parte, los valores del estadístico y el valor crítico, al 5%, de la prueba Bootstrap, son 1.8093 y 4.2789, por lo que tampoco se rechaza H_0. Lo mismo ocurre cuando se aplica esta metodología a un conjunto de datos cuyos valores del estadístico y del valor crítico conducen al rechazo de la hipótesis nula de no preferencia.

level. The procedure consists in: 1) From the original data of the sample, the estimators $\hat{\theta}$ and $\hat{\pi}$ of the Beta-Binomial model are obtained using the approximation of Newton-Raphson where $\pi=\alpha/(\alpha+\beta)$ and $\theta=1/(\alpha+\beta)$. 2) In a manner similar to step 1), the estimator of parameter θ under H_0 and $\hat{\theta}_{H_0}$ are obtained, and with both results the likelihood ratio statistic $\hat{\delta}=-2\log\hat{\lambda}$ is calculated. 3) 1000 Bootstrap samples are taken from the Beta-Binomial distribution under H_0: $\pi=\frac{1}{2}$ and $\hat{\theta}_{H_0}$. Each re-sampling must be the same size as the initial sample. 4) The statistic $\hat{\delta}$ is calculated for each Bootstrap sample, and with these the empirical distribution function $\hat{\delta}$ is constructed. 5) The non-preference test of hypothesis, H_0: $\pi=\frac{1}{2}$, with α=0.05, is equivalent to obtaining the 95.0, $q_{0.95}$, percentile of the Bootstrap percentile confidence interval, and to reject H_0 if $\hat{\delta}>q_{0.95}$.

Results and Discussion

To illustrate this methodology, a program with SAS' IML[3] was constructed, and was run with an n=20-sized sample for each of m=10 binomials, using the data 8, 19, 13, 1, 7, 20, 15, 13, 8, 20, and a total of B=1000 Bootstrap samples. The χ^2 and Fitzmaurice methods were applied to detect over-dispersion. From the results it was concluded that the phenomenon of over-dispersion is present in the data of the example (Méndez and Ramírez, 2001).

With these results the application of the proposed methodology is quite rational. The program has its fundament in the equivalence between the test and the Bootstrap percentile interval. Thus, obtaining the 95.0 percentile of the empirical distribution of the statistic $\hat{\delta}$ is equivalent to rejecting H_0 if $\hat{\delta}>q_{0.95}$. The statistic and the 5% critical value of the likelihood ratio test are $\hat{\delta}$ =1.8093 and 3.8414, respectively, and therefore the null hypothesis of non-preference of the insects is not rejected. On the other hand, the values of the statistic and the 5% critical value of the Bootstrap test are 1.8093 and 4.2789, and therefore H_0 is not rejected either. The same occurs when this methodology is applied to a set of data whose statistic and critical values lead to the rejection of the null hypothesis of non-preference.

Conclusions

It can be concluded that, when conducting experiments with dichotomous variables and over-dispersion exists in the set of data under study, the Bootstrap methodology, using the likelihood ratio of the Beta-Binomial distribution is a good option for performing the statistical analysis, as it adequately approximates the exact distribution.

—End of the English version—

[3] Disponible en anexo en ❖ Available as an annex in: http://www.colpos.mx/agrocien/agrociencia.htm/bimestral/2002/may-jun-02/boots.sas

Conclusiones

Se puede concluir que cuando se experimenta con variables dicotómicas y en el conjunto de datos bajo estudio está presente la sobredispersión, la metodología Bootstrap, usando la razón de verosimilitud de la distribución Beta-Binomial, es una buena opción para realizar el análisis estadístico, en el sentido que aproxima adecuadamente la distribución exacta.

Literatura Citada

Babu, G. J., and K. Singh. 1983. Inference on means using the bootstrap. Annals of Statistics. 11, 338-370.

Bailey, N. T. J. 1957. The Mathematical Theory of Epidemics. Griffin, London.

Bickel, O. J., and D. A. Freedman. 1981. Some asymptotic theory for the bootstrap. Annals of Statistics. 9, 1196-1217.

Efron, B.1979. Bootstrap Methods. Another look at the Jackknife. Annals of Statistics 7, 1-26.

Efron, B. 1982. The Jackknife, the Bootstrap and other Resampling Plans. Philadelphia. Society for Industrial and Applied Mathematics. 52 p.

Méndez O., C. y G. Ramírez V. 2001. Prueba de razón de verosimilitud para el estudio de preferencias con dos opciones. Comparación del modelo Beta-binomial con métodos alternativos para el estudio de preferencias sobre dos opciones. Agrociencia, Vol. 35, No. 5: 543-550.

Vaillant, J., and S. Derridj. 1992. Statistical Analysis of Insect preference in Two-Choice Experiments. Journal of Insect Behavior, Vol. 5, No. 6, pp: 211-230.