

# Estimación de proporciones multinomiales

Jairo Alfonso Clavijo Méndez  
Universidad del Tolima  
Abril de 2005

En esta conferencia haremos algunas consideraciones acerca del tamaño de muestra necesario para estimar proporciones bajo muestreo aleatorio simple. En particular el tema se centrará en el caso de distribuciones multinomiales, el más utilizado en la práctica y, a la vez, el más desconocido.

Comenzaremos recordando algunos conceptos básicos como son las distribuciones binomial y multinomial y la fórmula para el cálculo del tamaño de muestra para proporciones binomiales.

## 1. La distribución binomial:

Supóngase que  $p$  es un valor en el intervalo  $[0, 1]$ , el cual puede ser interpretado como la probabilidad con la cual se da un éxito en un experimento de Bernoulli (cualquier experimento que tenga sólo dos resultados, *éxito* y *fracaso*, recibe este nombre). Supóngase ahora que el experimento se repite  $n$  veces en forma independiente (es decir, el resultado obtenido en un determinado momento no depende de los resultados obtenidos anteriormente). Si  $p$  es la probabilidad de éxito en una realización del experimento entonces  $q = 1 - p$  es la probabilidad de fracaso.

A partir de la conocida fórmula para calcular una potencia de cualquier binomio, se obtiene:

$$(p + q)^n = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \quad \text{donde} \quad \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Puesto que  $p + q = 1$  se concluye que la suma anterior vale 1. Esto permite definir una función de densidad, mediante la fórmula:

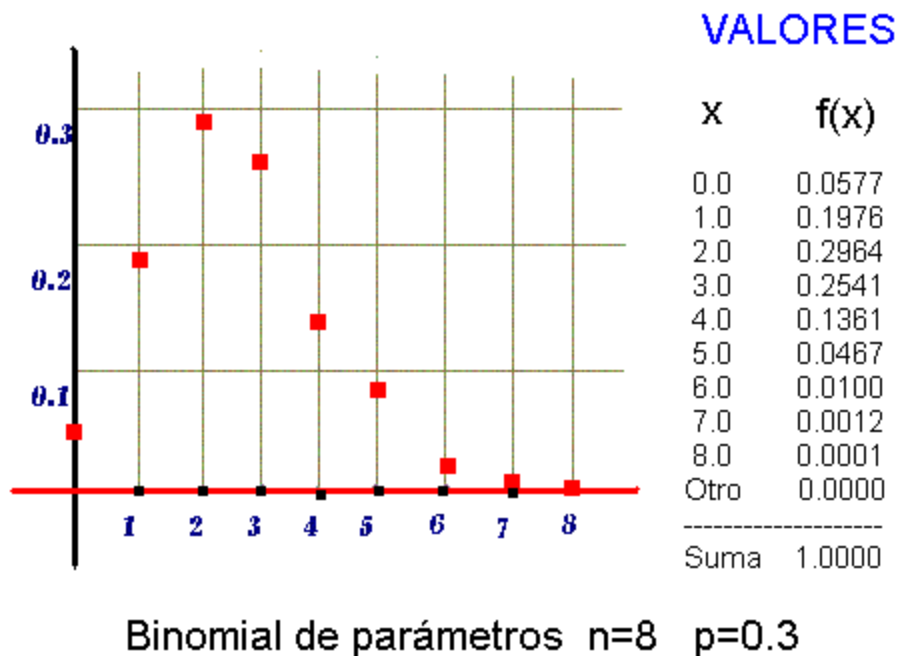
$$f(x) = \begin{cases} \binom{n}{x} p^x q^{n-x} & \text{Si } x = 0, 1, 2, \dots, n \\ 0 & \text{En cualquier otra parte} \end{cases}$$

Es claro que la función anterior puede ser interpretada como la función que mide la probabilidad de que se den  $x$  éxitos en las  $n$  repeticiones del experimento de Bernoulli.

La función de densidad definida según la fórmula anterior, tiene como propiedad interesante el ser simétrica en el caso en que  $p = q = 0.5$  y ser asimétrica en los demás casos. Pero, al ser intercambiables los papeles de  $p$  y  $q$ , cada caso de asimetría izquierda tiene una imagen especular de asimetría derecha. Cada par de valores  $n$  y  $p$  da origen a una distribución de probabilidad, denominada **binomial de parámetros  $n$  y  $p$** , comúnmente simbolizada como  $b(n,p)$

Si se conoce  $p$  se pueden calcular los valores de  $f(x)$  y los de su acumulada (función de distribución), definida como  $F(x) = \sum_{t \leq x} f(t)$ . Estos valores corresponden respectivamente a la probabilidad de que ocurran exactamente  $x$  éxitos en los  $n$  ensayos y la probabilidad de que el número de éxitos sea a lo más  $x$ .

La gráfica 1 corresponde a la función de densidad binomial con  $p = 0.3$  y  $n = 8$ . Como se ve, ella presenta una asimetría de tipo positivo, correspondiendo su máximo al caso  $x = 2$  éxitos, cuya probabilidad es 0.2964.



Gráfica 1. Una distribución binomial

Si una variable aleatoria discreta  $X$  tiene distribución binomial de parámetros  $n$  y  $p$ , se cumple  $E(X) = np$  y  $V(X) = np(1-p) = npq$  donde  $q = 1-p$  es la probabilidad de fracaso en cada experimento.

## 2. Distribución multinomial

La inmediata generalización de las variables binomiales que miden el número de éxitos y, por tanto de fracasos, en  $n$  experimentos de dos resultados, son las variables multinomiales que miden el número de ocurrencias de cada resultado en  $n$  experimentos diferentes cada uno con  $m$  posibles resultados (categorías). Un ejemplo de tales experimentos es la observación de la luz de un semáforo en funcionamiento. El experimento tiene tres posibles resultados a saber: amarillo, rojo, verde (A,R,V). El semáforo siempre estará en alguno de estos tres estados con ciertas probabilidades, digamos  $p_A, p_R, p_V$ , tales que  $p_A + p_R + p_V = 1$ . Una variable aleatoria -trinomial en este caso- contará el número de veces  $n_A$  que el semáforo esté en amarillo, el número de veces  $n_R$  que esté en rojo y el número de veces  $n_V$  en que se encuentre en verde, al ser observado  $n$  veces. Por supuesto  $n_A + n_R + n_V = n$ .

La función de densidad para la distribución de una variable aleatoria  $X$  multinomial depende de los parámetros  $n, p_1, p_2, \dots, p_m$  donde  $p_i$  es la probabilidad de que el resultado del experimento se encuentre en la  $i$ -ésima categoría, con  $i = 1, 2, \dots, m$  y

$$\sum_{i=1}^m p_i = 1. \text{ Se tiene en tal caso } f(x_1, x_2, \dots, x_m) = \frac{n!}{x_1! x_2! \dots x_m!} p_1^{x_1} p_2^{x_2} \dots p_m^{x_m} \text{ con } \sum_{i=1}^m x_i = n.$$

En este caso es posible considerar  $m$  variables aleatorias  $X_i$ , cada una de las cuales cuenta la cantidad de resultados que clasifican en la  $i$ -ésima categoría. Es claro que entonces  $X_i \sim b(n, p_i)$  y, por tanto,  $E(X_i) = np_i$ ,  $V(X_i) = np_i(1-p_i)$ , sin embargo, tales variables no son independientes pues siempre estarán ligadas por la restricción  $\sum_{i=1}^m x_i = n$ .

El ejemplo de variable multinomial más interesante para esta conferencia corresponde a las preguntas de una encuesta en las cuales hay más de dos alternativas de respuesta. Claramente, si se hace una pregunta con  $m$  posibles respuestas exhaustivas y mutuamente excluyentes, el encuestado escoge solo una alternativa de respuesta y, en consecuencia, es igual que realizar un experimento de  $m$  posibles resultados de los cuales se observa solamente uno de ellos.

Los  $n$  diferentes formularios de una encuesta en la que se contesta una pregunta de tipo multinomial pueden verse como una muestra aleatoria de tamaño  $n$  con la cual se estiman las probabilidades  $p_1, p_2, \dots, p_m$  para cada categoría de la pregunta. Volveremos sobre este tema más adelante.

### 3 Estimación de una proporción binomial

Consideremos inicialmente el caso de una variable aleatoria binomial, la cual representa una población de tamaño  $N$  dividida en dos clases (por ejemplo, poblaciones animales divididas en machos y hembras). Estas clase se denotarán  $A$  y  $A'$ . Se define la proporción de  $A$  como el número  $P = \frac{A}{N}$  donde  $A$  es el número de individuos en la clase  $A$ . Nótese que la proporción multiplicada por 100 es igual al porcentaje de individuos que se encuentran en la clase  $A$  y que  $P$  es la probabilidad de que al seleccionar aleatoriamente un elemento de la población, dicho elemento pertenezca a la categoría  $A$ . Por supuesto que  $Q = 1 - P = \frac{A'}{N}$  donde  $A'$  es el número de elementos que hay en la categoría  $A'$  de la población.

Consideremos una variable de Bernoulli definida por  $X = \begin{cases} 1 & \text{si } A \\ 0 & \text{si } A' \end{cases}$  Esta variable anota un éxito si el elemento seleccionado es de la clase  $A$  y un fracaso si es de  $A'$ . Consideremos ahora una muestra aleatoria  $\{X_1, X_2, \dots, X_n\}$  de tales variables y sea  $S$  su suma, esto es:  $S = X_1 + X_2 + \dots + X_n$ . Se tiene entonces que  $S$  tiene distribución binomial de parámetros  $P$  y  $n$ . De aquí que  $E(S) = nP$  y  $V(S) = nPQ$ .

Se cumple  $\bar{X} = \frac{1}{n}S$  y de aquí  $E(\bar{X}) = \frac{1}{n}E(S) = P$  y  $V(\bar{X}) = \frac{1}{n^2}V(S) = \frac{PQ}{n}$

Lo anterior sugiere utilizar  $\bar{X}$  como estimador insesgado de  $P$ . Resulta evidente, sin embargo, que  $\bar{X} = \frac{a}{n}$  donde  $a$  es el número de elementos de la clase  $A$  que aparecen en la muestra. Utilizaremos la expresión  $p = \frac{a}{n}$  para el estimador de  $P$ . Con lo dicho anteriormente, el estimador propuesto es insesgado.

La construcción de intervalos de confianza para  $P$  presenta problemas de tipo teórico no fáciles de resolver debido a que se debe tener en cuenta la distribución del estimador

$\bar{X} = \frac{a}{n}$ , distribución que no es fácil de determinar ya que corresponde a la distribución de una combinación lineal de variables aleatorias binomiales. Por esta razón en la práctica se utiliza una aproximación normal a la binomial, mediante la variable aleatoria  $Z = \frac{S - E(S)}{\sqrt{V(S)}} = \frac{S - nP}{\sqrt{nPQ}}$ , aproximación es que válida siempre que  $n$  sea grande.

En la expresión anterior se puede dividir numerador y denominador entre  $n$  lo que produce

$Z = \frac{\bar{X} - P}{\sqrt{\frac{PQ}{n}}} = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$  variable que, tendrá una distribución asintótica normal

estándar. En otras palabras,  $p$  se puede considerar como un estadístico con distribución normal de media  $P$  y varianza  $\frac{PQ}{n}$  siempre que  $n$  sea grande. Esto es:  $p \sim N(P, \frac{PQ}{n})$

lo que nos permite construir intervalos de confianza para la proporción  $P$  mediante la fórmula:  $(p - z_{\alpha/2} \sqrt{\frac{PQ}{n}}, p + z_{\alpha/2} \sqrt{\frac{PQ}{n}})$

La deducción anterior se hace bajo el supuesto de que la población es infinita. Sin embargo en la práctica se deben hacer dos correcciones: una por tratarse de poblaciones finitas y la otra es una corrección por continuidad debido a que una distribución discreta (la binomial) se está aproximando por una continua (la normal). De esta manera la fórmula para los intervalos de confianza de la proporción es realmente:

$$(p - z_{\alpha/2} \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{PQ}{n}} - \frac{1}{2n}, p + z_{\alpha/2} \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{PQ}{n}} + \frac{1}{2n})$$

Los valores  $P$  y  $Q=1-P$  necesarios para el cálculo del intervalo son valores poblacionales desconocidos. Más aún, se está utilizando el mismo valor  $P$  que se quiere estimar lo que parece un círculo vicioso. Por esta razón, en cambio de  $P$  se utiliza la estimación suya obtenida con la muestra, pero este cambio altera ligeramente la fórmula, de la siguiente manera:

$$(p - z_{\alpha/2} \sqrt{\frac{N-n}{N}} \sqrt{\frac{pq}{n-1}} - \frac{1}{2n}, p + z_{\alpha/2} \sqrt{\frac{N-n}{N}} \sqrt{\frac{pq}{n-1}} + \frac{1}{2n})$$

Expresión que en la práctica se usa para estimar una proporción binomial, es decir de dos categorías, con muestras grandes en poblaciones finitas.

Si se trata de estimar proporciones binomiales en poblaciones infinitas desaparece el factor de corrección por finitud,  $\sqrt{\frac{N-n}{N-1}}$  ya que éste puede considerarse igual a 1.

Podemos ahora calcular el tamaño mínimo de muestra necesario para hacer una estimación de una proporción binomial.

En primer lugar, en una población infinita, se tiene  $e = z_{\alpha/2} \sqrt{\frac{PQ}{n}}$  de donde se deduce  $n = \frac{z^2 PQ}{e^2}$ , expresión que comúnmente se denota:  $n_{\infty} = \frac{z^2 PQ}{e^2}$

Para el cálculo de este valor es necesario tener conocimiento de cuál puede ser el valor de  $P$ , lo que parece un círculo vicioso pues precisamente  $P$  y  $Q$  se van a estimar. Teniendo en cuenta que la varianza del estimador  $p$  es  $V(p) = npq = n(p - p^2)$  se concluye que dicha varianza es máxima cuando  $p = q = \frac{1}{2}$ . En consecuencia,  $n_{\infty} = \frac{z^2}{4e^2}$  es un valor que garantiza una muestra suficiente para la estimación de  $P$ . Esta situación,  $p = q = \frac{1}{2}$ , corresponde al **peor caso** pues exagera un poco el tamaño de muestra debido a que corresponde a la situación de máxima varianza.

Cuando la población es finita, de tamaño  $N$ , se tiene  $e = z_{\alpha/2} \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{PQ}{n}}$  o, lo que es equivalente:  $n(N-1)e^2 = (N-n)z^2 PQ$  de donde se concluye que  $n = \frac{Nz^2 PQ}{(N-1)e^2 + z^2 PQ}$

y, dividiendo numerador y denominador por  $Ne^2$ , se obtiene finalmente la expresión:

$$n = \frac{n_{\infty}}{1 + \frac{n_{\infty} - 1}{N}}$$

Esta es la fórmula práctica para el cálculo del tamaño de muestra bajo M.A.S: primero se calcula  $n_{\infty}$  y luego, si es necesario, se corrige para poblaciones finitas.

#### 4. Estimación y tamaño de muestra para los valores de probabilidad en distribuciones multinomiales

Como se dijo antes, si  $X$  tiene distribución multinomial con  $m$  categorías, se puede tomar una muestra aleatoria de tamaño  $n$  en la población definida por  $X$  y, de una manera similar al caso divariado, tomar  $p_i = \frac{a_i}{n}$  como el estimador de  $P_i$ , la probabilidad de la categoría  $A_i$  donde  $a_i$  es el número de elementos de la categoría  $A_i$  presentes en la muestra.

Las estimaciones puntuales así obtenidas son correctas pero no resulta sencillo construir intervalos de confianza para ellas, ya que, como se vió antes, aunque las variables  $X_i$  que hacen conteos por categoría son binomiales y, en consecuencia, la distribución de cada  $p_i$  podría aproximarse por una normal, como se hizo en el caso binomial, las variables  $X_i$  no son independientes y, por tanto, tampoco lo serán los estimadores  $p_i$ , los que están ligados por la restricción  $\sum_{i=1}^m p_i = 1$ . Esto hace que la construcción de un intervalo de confianza para  $p_i$  afecte los intervalos de confianza para los otros  $p_j$ . Más complejo aún es determinar el tamaño de muestra mínimo para estimar las proporciones de las diferentes categorías.

Una aproximación al problema es agrupar las categorías en dos grupos: de una parte considerar la categoría de interés para la que se desea estimar la proporción y de otra reunir las categorías restantes en una sola categoría con lo cual el problema ha sido forzado a parecerse a una situación binomial y aplicar luego la teoría de vista anteriormente para variables binomiales. Esto, por supuesto, no es más que una salida de emergencia pues no siempre proporciona una correcta solución, primero porque no permite construir intervalos de confianza para cada  $P_i$  y, segundo, porque exigiría un tamaño de muestra calculado específicamente para la proporción de interés, lo que usualmente no se hace.

Mucho más realista, aunque también limitado, sería considerar una situación tipo Bonferroni, donde se construyan intervalos simultáneos para todas las probabilidades  $P_i$ . Es decir, dado  $\alpha$ , obtener un conjunto de  $m$  intervalos  $J_i$  para los cuales se cumpla que

$$P((p_1 \in J_1) \wedge (p_2 \in J_2) \wedge \cdots \wedge (p_m \in J_m)) = 1 - \alpha$$

Los intervalos que cumplen las condiciones anteriores se obtienen solucionando un complejo sistema de ecuaciones del tipo:

$$p_i = \frac{(\mathbf{c}^2 + 2n_i \pm \mathbf{c}^2(\mathbf{c}^2 + 4n_i(N - n_i)/N)^{\frac{1}{2}}}{2(N + \mathbf{c}^2)}$$
 donde  $\mathbf{c} \sim \mathbf{c}_{m-i,a}^2$  y los  $p_i$  son las probabilidades verdaderas de cada categoría ver [\*]

Los valores  $p_i$  son desconocidos, así que con frecuencia se tomen iguales (esto ya introduce incorrecciones!) además que funciona mejor para pequeños valores de  $m$  (cuando mucho 4)

Ante las anteriores dificultades se han propuesto varias soluciones empíricas, muchas de ellas basadas en la propuesta de Cochran de agrupar las modalidades en dos grupos y hacer un tratamiento binomial. Por ejemplo, Yarnold(1970), basándose en estudios de simulación, proponía que el tamaño de muestra para poder aplicar esta metodología debería satisfacer  $np_i \geq 5q$  para todo  $i = 1, 2, \dots, m$  con  $m \geq 3$  siendo  $q$  la proporción de categorías para las cuales  $np_i < 5$ .

Ya en 1964 Queensbury y Hurst presentaron un método de construcción simultánea de intervalos basándose en la distribución Ji cuadrado aproximada de la suma de valores observados menos valores estimados al cuadrado dividida entre valores estimados. Goodman en 1965 construyó intervalos más cortos basándose en aproximaciones normales y la desigualdad de Bonferroni para imponer una cota a la probabilidad de que todos los intervalos fuesen simultáneamente correctos.. En 1974 Angers, basándose en el método de Goodman presentó un método gráfico para fijar el tamaño de muestra usando valores conocidos “a priori” de los parámetros. Tortora en 1978, introduce la idea del **peor caso** para distribuciones multinomiales, por analogía con el peor caso de proporciones binomiales. El método de Tortora fue criticado por Angers quien estableció que dicho método era más conservativo de lo necesario y propuso revisarlo usando un valor de 0.5 para cada parámetro. Este método es computacionalmente tedioso en general pero puede simplificarse si se suponen intervalos de igual longitud para todos los parámetros . Este último supuesto ha sido adoptado en casi todas las situaciones para estimar el tamaño de muestra, siguiendo casi siempre el método propuesto por Cochran.

Presentaremos entonces el método propuesto ya mencionado que ha sido revisado por S Thompson, con el supuesto de que todos los intervalos son de igual longitud y usando el concepto de *peor caso* (máxima varianza) en distribuciones multinomiales como una generalización del *peor caso* en distribuciones binomiales. En la práctica no existen razones fuertes para creer que algunos intervalos sean más cortos que otros.

El objetivo es encontrar el menor tamaño de muestra  $n$  para una muestra aleatoria extraída de una población multinomial de modo que la probabilidad de que todas las proporciones estimadas estén simultáneamente dentro de unas distancias especificadas de las verdaderas

proporciones sea al menos  $1-\alpha$ . Esto es:  $\Pr\left(\bigcap_{i=1}^m |p_i - \mathbf{p}_i| \leq d_i\right) \geq 1-\alpha$  donde  $\mathbf{p}_i$  es la proporción poblacional en la  $i$ -ésima categoría, mientras  $p_i$  es la proporción observada en dicha categoría. Asumiremos que la población es lo suficientemente grande como para poder ignorar las correcciones que se hacen por finitud al utilizar aproximación normal cuando se emplea muestreo aleatorio simple sin reemplazamiento.

El procedimiento general consiste en encontrar el  $n$  más pequeño tal que  $\sum_{i=1}^m \mathbf{a}_i \leq \alpha$  para todos los posibles valores del vector  $(\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m)$  siendo  $\mathbf{a}_i$  el nivel de significancia para cada parámetro, valores que, en principio, asumiremos iguales.

En la práctica lo anterior se logra mediante el siguiente algoritmo:

1. Para cada valor posible de los parámetros del vector escoja un valor de  $n$  y calcule  $\sum_{i=1}^m \mathbf{a}_i$  donde  $\mathbf{a}_i = 2(1-\Phi(z_i))$  siendo  $z_i = \frac{\sqrt{nd_i}}{\sqrt{\mathbf{p}_i(1-d_i)}}$ . Si  $\sum_{i=1}^m \mathbf{a}_i < \alpha$  entonces repita el procedimiento tomando un valor de  $n$  menor que el anterior. Por el contrario, si  $\sum_{i=1}^m \mathbf{a}_i > \alpha$  tome un valor de  $n$  mayor que el anterior.
2. Repita el paso anterior con todos los posibles valores que puede tomar el vector de parámetros para determinar el vector de parámetros  $\mathbf{p}_0$ , correspondiente al **peor caso**, el cual proporciona el mayor  $n$ . Tome este  $n$  como el tamaño de muestra.

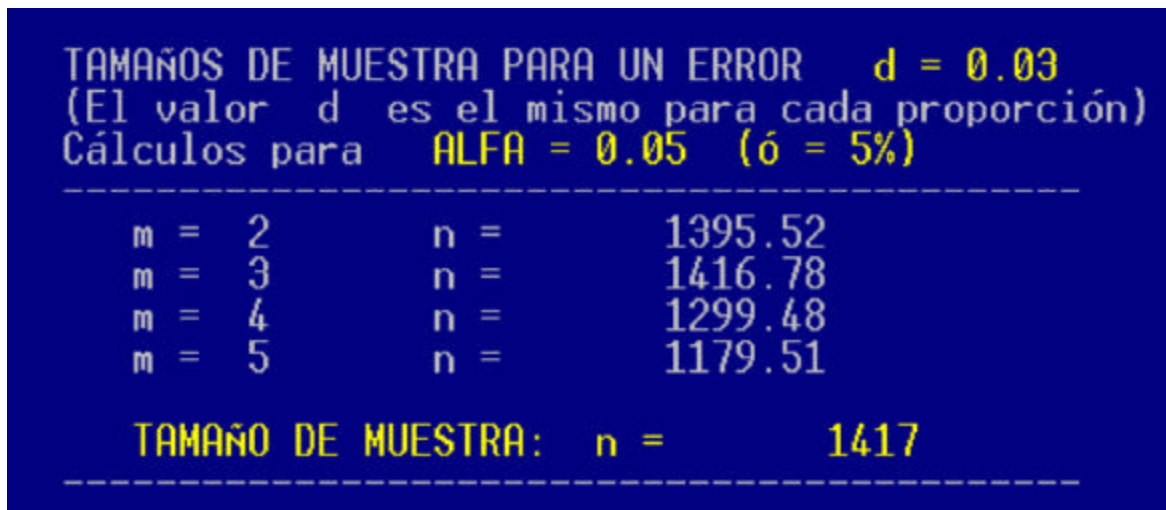
Puede verse que el peor caso para cada valor del vector de parámetros corresponde a la situación en que el correspondiente valor de parámetro es  $\frac{1}{m}$  mientras que los demás valores valen 0. Ante esto, el tamaño de muestra correspondiente estará dado por

$$n = \max_k \left\{ \frac{\left( z^2 \frac{1}{k} \left( 1 - \frac{1}{k} \right) \right)}{d^2} \right\} ..$$

siendo  $z$  el percentil superior correspondiente a  $100\frac{\alpha}{2k}\%$  bajo la normal estándar,  $d$  el valor común de todos los  $d_i$ , y  $k$  un entero menor o igual que el número  $m$  de categorías.

La versión 8.1 de ESM-plus incorpora una rutina para calcular el tamaño de muestra siguiendo el último procedimiento descrito para cualquier número de categorías entre 3 y

9, con  $\alpha = 0.05$  y errores de 0.01, 0.02, 0.03, ..., 0.09. La gráfica 2 muestra que para estimar proporciones con una variable multinomial de 5 categorías, con un error del 3% y un nivel de significancia del 95%, son necesarias 1417 observaciones bajo muestreo aleatorio simple.



Gráfica 2. Tamaño de muestra para proporciones multinomiales. ESM v8.1

## Referencias:

1. Clavijo M, Jairo A (2005); *Métodos Estadísticos*. Universidad del Tolima. Ibagué
2. Keeping, E.S (1995).; *Introduction to Statistical Inference*. Dover Publications. N.J. U.S.A
3. Quesenberry C y D.Hurst; *Large Simple Simultaneous Confidence Intervals for Multinomial Proportions*. Technometrics. Vol 6. No 2. Mayo 1964
4. Tortora, R.; *A note on Simple Size Estimation for Multinomial Populations*. The American Statistician. Vol 32 No 3. Agosto de 1978.
5. Eaton P.W.; *Yarnold's Criterion and Minimum Simple Size*. The American Statistician. Vol 32 No 3. Agosto de 1978.
6. Thompson, S.; *Simple Size for Estimating Multinomial Proportions*. The American Statistician. Vol 41 No 1. Febrero de 1987