TAMAÑO DE MUESTRA PARA POBLACIONES MULTINOMIALES EN MUESTREO BIETÁPICO

Svetlana Ivanovna Rudnykh.
Departamento de Física Universidad del Atlántico Km 7 antigua vía a Puerto Colombia, A.A. 1890, Barranquilla, Colombia, Svetarudn@hotmail.com
Material Tesis de Especialización en Estadística, Convenio Universidad Nacional - Universidad del Atlántico

Resumen. En esta investigación se presenta un algoritmo que permite establecer un tamaño de muestra para poblaciones multinomiales que requieren el mínimo costo para realizar un muestreo bietápico.
Palabras clave: distribución binomial, muestreo bietápico, tamaño de muestra, distribución multinomial.

Abstract. In this research, I introduce an algorithm that allows establishing an optimal sample size for multinomial populations in two-step survey.
Key-Words: binomial distribution, two-step survey, sample size, multinomial distribution.

1. Introducción

El cálculo del tamaño de muestra para estimar parámetros de proporciones con distribución multinomial se ha convertido en una tarea cotidiana en las investigaciones sociales (ver MEDINA 1998). Este problema ha sido analizado, entre otros, por COCHRAN (1963), TORTORA (1978), THOMPSON (1987) y ANGERS (1974, 1984), quienes han podido aplicar sus métodos en actividades tan disímiles como control de calidad, opinión pública, antropología, teoría del juego, biología y estudios de simulación.
En la estimación del número de unidades que forman parte de una muestra, debe ser considerada la varianza de la variable de interés, así como la precisión con la que se desean obtener las estimaciones y la confianza requerida, los dominios de estudio y el esquema de muestreo. Sin embargo, la forma de abordar este problema es muy compleja y la teoría conocida hasta hoy presenta soluciones muy puntuales a casos particulares.
En este estudio se consideró un procedimiento de muestreo formado por dos etapas (bietápico). Se supuso, además, que en la etapa inicial se extrae una muestra aleatoria simple sin reemplazo de n UPM (Unidades Primarias de Muestreo, las cuales están conformadas a su vez por unidades de menor tamaño), de un total de N que componen la población objetivo, en una segunda etapa se extrae una muestra aleatoria simple sin reemplazo de m USM (Unidades Secundarias de Muestreo) de las M que componen cada UPM. En otras palabras, el procedimiento bietápico aquí considerado es aplicado a una población de N UPM, en donde cada UPM tiene igual tamaño M. De aquí son extraídas m USM para

ser examinadas y estimar la proporción de una característica de interés (esta variable es de tipo multinomial).
Bajo estas condiciones, se propone un algoritmo para estimar los tamaños de muestra (n, m) de poblaciones multinomiales en el muestreo bietápico.

2. Tamaños de Muestra de Poblaciones Binomiales en Muestreo Bietápico

El procedimiento usual para el cálculo del tamaño de muestra en el esquema bietápico cuando se estima el parámetro desconocido (P) de poblaciones binomiales implica optimizar una función de costo teniendo en cuenta las restricciones contenidas en la expresión que se obtenga de la varianza del estimador $(\widehat{P})$.

P se define aquí como la proporción poblacional en la i- ésima UPM, o también la razón entre el número total de unidades en la i-ésima UPM que posee la característica de interés y M (tamaño de cada UPM).

Por su parte $\widehat{P} = \dfrac{1}{n}\sum_{i=1}^{n}\widehat{P}_i$ y $\widehat{P} = \dfrac{a_i}{m}$, siendo a_i el total de unidades USM que poseen atributo de interés y pertenecen a la UPM i.
Aplicando el teorema de Madow (PÉREZ 2000) se obtiene la varianza del estimador $(\widehat{P})$, cuya expresión es la siguiente:

$$V\left(\widehat{P}\right) = (1 - f_1)\frac{S_b^2}{n} + (1 - f_2)\frac{S_w^2}{nm} \tag{1}$$

Donde $f_1 = \dfrac{n}{N}$, $S_b^2 = \dfrac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2}{(N-1)}$, $f_2 = \dfrac{m}{M}$ y $S_w^2 = \dfrac{\sum_{i=1}^{N} MP_i\left(1 - P_i\right)}{N(M-1)}$. Una función de costo del muestreo dependiente de n y m se puede presentar de la siguiente forma:

$$C = nC_1 + nmC_2, \tag{2}$$

donde C_1 y C_2 son los costos de muestreo correspondientes a cada unidad primaria y secundaria en la encuesta, respectivamente.
Utilizando la metodología de Lagrange se logra optimizar (minimizar) la función de costos (2) bajo las condiciones de (1) y de esa forma encontrar m y n óptimos:

$$m_{opt.} = \sqrt{\frac{C_1 M S_w^2}{C_2(MS_b^2 - S_w^2}}$$

$$n = \frac{C}{C_1 + mC_2}.$$

Se observa que el tamaño óptimo de m aumenta proporcionalmente a $\sqrt{\frac{C_1}{C_2}}$, pero no es muy sensible a pequeños cambios en $\frac{C_1}{C_2}$, también se observa que $m_{opt.}$ no depende de C ni de n.

3. Tamaño de Muestra de Poblaciones Multinomiales en Muestreo Monoetápico

Para el cálculo del tamaño de muestra en distribuciones multinomiales se han desarrollado procedimientos que se orientan a resolver parte de los problemas teóricos que comprende la estimación simultánea de los parámetros en poblaciones multinomiales, pero no logran optimizar sus soluciones para todos los casos.

Así, en Angers (1984) se propone un método que consiste en elegir de manera arbitraria un tamaño de muestra n y calcule los k cocientes por medio de la expresión: $nd_i^2 P_i(1-P_i)$; $i = 1, 2, ..., k$ (categorías de la variable de diseño) que representan los valores de las abscisas, mientras que en el eje de las ordenadas se ubican los niveles de confianza ($0 \leq \alpha \leq 0{,}10$) y ($0 \leq \alpha \leq 0{,}01$).

Posteriormente, se buscan en las gráficas propuestas los valores obtenidos en el eje de las abscisas, a fin de identificar los correspondientes niveles de confianza ($\alpha_i\,'s$); y se compara la $\sum \alpha_i$ con el valor de definido por el investigador. El criterio que se utiliza para decidir es que si la sumatoria $\left(\sum \alpha_i\right)$es mayor (menor) que α, entonces el tamaño de muestra propuesto es muy pequeño (grande), por lo que se deberá modificar el tamaño de muestra en múltiplos de n y continuar con el procedimiento descrito hasta encontrar un intervalo ($n_1 \leq n \leq n_2$) que contenga el valor buscado. Cuando se logre ubicar el intervalo, el número final de observaciones se obtiene por medio de interpolación lineal simple.

Este procedimiento permite obtener el límite empírico para el tamaño de muestra cuando se asume que los intervalos de confianza tienen amplitudes iguales y no se hacen restricciones a los aparte de que $\sum \alpha_i = \alpha$.

En la propuesta de Angers (*ut supra*), se nota que el tamaño de muestra se incrementa con el aumento del número de categorías k. Este es un resultado irregularmente conservativo, puesto que falla al tomar en cuenta la restricción $\left(\sum_{i=1}^{k} P_i = 1\right)$de los parámetros multinomiales.

Para la determinación del tamaño de muestra, se requiere definir la precisión de cada parámetro de la distribución multinomial. Esta situación representa una diferencia sustantiva respecto del procedimiento tradicional, en donde generalmente se elige una variable de diseño y sobre ella se determina el número de observaciones necesarias para realizar la investigación. De esta manera, suponga que se desea una precisión absoluta para cada celda; entonces, se tiene que:

$$P_i - \delta_i = P_i - \sqrt{\frac{\chi^2_{(1,\alpha/k)} P_i(1-P_i)}{n}},$$
$$P_i + \delta_i = P_i + \sqrt{\frac{\chi^2_{(1,\alpha/k)} P_i(1-P_i)}{n}},$$

Despejando el valor de δ_i en las anteriores expresiones, se obtiene:

$$\delta_i = \sqrt{\frac{\chi^2_{(1,\alpha/k)} P_i(1-P_i)}{n}},$$

y resolviendo para n se encuentra que el tamaño de muestra necesario para estimar cada celda con una precisión δ_i es:

$$n = \max_i \frac{\chi^2_{(1,\alpha/k)} P_i(1-P_i)}{\delta_i^2}.$$

En 1987, Thompson plantea que el método propuesto por Angers en 1984 era el óptimo de los procedimientos existentes, pero resultaba muy tedioso en su aplicación, por lo que propuso una manera de determinar el "peor de los casos" (worst case $P_i = 0{,}5$) para un vector de parámetros multinomiales cuando se desean obtener intervalos de confianza simultáneos para cada uno de los componentes del vector P.

Thompson (*ibid*) plantea que el objetivo consiste en determinar el tamaño de muestra n para una variable aleatoria de una distribución multinomial, de tal forma que la probabilidad de que todas las proporciones estimadas de manera simultánea estén contenidas en el intervalo sea menor que $(1-\alpha_i)$, esto es,

$$P_r\left\{\bigcap_{i=1}^{k} |p_i - P_i| \leq \delta_i\right\} \geq 1-\alpha,$$

en donde P_i es la proporción de observaciones en la i-ésima categoría en la población, p_i la proporción observada en la muestra, k el número de categorías y

$$\alpha_i = P_r\left\{|Z_i| \geq \frac{\delta_i\sqrt{n}}{\sqrt{P_i(1-P_i)}}\right\} = 2(1-\phi(Z_i)),$$

en donde Z_i es la variable normal estandarizada, ϕ la función acumulativa de probabilidad y

$$Z_i = \frac{\delta_i\sqrt{n}}{\sqrt{P_i(1-P_i)}}.$$

Cuando $k = 2$ y $\delta_i = \delta_1 = \delta_2$ se trata de una distribución binomial y el tamaño de muestra se determina de la manera tradicional:

$$n = \frac{Z^2 P_i(1-P_i)}{\delta_i^2}$$

Si la proporción P_i es desconocida, se utiliza el criterio de máxima varianza (worst case) con $P_i = 0{,}5$.

4. Efecto de Categorías para la Determinación de losTamaños de Muestra

Para tratar de resolver la situación que se presenta en el cálculo de tamaño de muestra en encuestas complejas, en donde no se cuenta con una fórmula para la varianza de las proporciones de las categorías de la variable de interés, siguiendo la propuesta de Kish (1972) al definir un factor de ajuste que a partir de una muestra aleatoria simple permite aproximarse al número de selecciones necesarias para un diseño de conglomerados, proporciona la misma varianza, se define aquí el tamaño efectivo de muestra como,

$$n_e = n_0 * efdk$$

donde n_0 es el tamaño de muestra obtenido según el procedimiento clásico de Cochran, $efdk$ es el efecto de diseño que, en la situación aquí analizada, sería el efecto de k categorías. Este efecto se expresa como las variaciones de los tamaños de muestra propuestos por los distintos procedimientos entre el tamaño de muestra de aproximación clásica (COCHRAN 1963) (ver tabla 1).
El tamaño efectivo calculado puede ser interpretado como la cantidad de información contenida en una muestra multinomial.
La subvaloración de la aproximación clásica (COCHRAN 1963) en el caso de la estimación de proporciones para poblaciones multinomiales con más de 2 categorías y por ende en la determinación del tamaño de muestra, se debe a la consideración no realista de que todos los parámetros son iguales a 0.5 (peor de los casos) y que la suma de los mismos es igual a 1.
Si en lugar de considerar poblaciones binomiales se consideran ahora poblaciones multinomiales, en otras palabras, si en lugar de estimar una sola proporción interesa estimar k proporciones de categorías de una variable, la varianza dentro de las unidades secundarias aumenta. Este aumento de la varianza se debe a la estimación simultánea de k proporciones de la variable.
Por otra parte, al expresar la varianza del estimador de la proporción en el modelo binomial mediante el coeficiente de correlación intraconglomerados, se observa que dicha varianza es igual al producto de varianza del estimador de la proporción en el muestreo aleatorio simple cuando el tamaño de muestra es mn por el factor $(1 + (m - 1)\rho)$, que es llamado por Kish (1972) efecto de diseño.
En resumen, el efecto de diseño dado por la razón entre la varianza del estimador de la proporción para el muestreo en etapas y varianza del estimador bajo el muestreo aleatorio simple depende vitalmente de m, el tamaño de muestra de unidades secundarias, y no tanto del tamaño de muestra de unidades primarias n.
En el cambio de poblaciones binomiales a poblaciones multinomiales siguiendo un esquema bietápico el tamaño de muestra que resulta incrementado sustancialmente es el de las unidades secundarias (m), y su influencia se puede medir por el efecto de categorías $efdk$ expresada de la siguiente manera:

$$m = m' * efdk,$$

donde es el tamaño de muestra de unidades secundarias y

$$efdk = \frac{\widetilde{n}}{\widetilde{n}'},$$

donde $\widetilde{n}'$es el tamaño de muestra propuesto por Cochran (1963) de la aproximación clásica y $\widetilde{n}$ es el tamaño de muestra propuesto por distintos autores para poblaciones multinomiales.

5. Algoritmo para el Tamaño de Muestra

Los tamaños de muestra de una población multinomial para el muestreo bietápico pueden encontrarse de una manera práctica desarrollando el siguiente procedimiento:

- Se obtienen los tamaños de muestra n de unidades primarias y m' (número de unidades secundarias) para poblaciones binomiales en muestreo bietápico.
- Escoger el procedimiento de aproximación de estimadores que más se ajusta a situación planteada en el problema que se resuelve (Tortora 1978, Angers 1984 o Thompson 1987).
- Luego que se ha escogido el procedimiento, ir a la celda correspondiente en la tabla 1, de acuerdo a los valores de α y k, y localizar el valor del efecto de categoría para este caso.
- Multiplicar m' por valor del efecto de categoría hallado en paso anterior. El resultado es aproximadamente el tamaño de muestra de las unidades secundarias recomendado por el procedimiento escogido.
- El tamaño de muestra n de unidades primarias fue el obtenido con la aproximación binomial al inicio del algoritmo.

Conf									Ang	Thom
α	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	1984	1987
0.0001	1.14	1.18	1.20	1.22	1.24	1.26	1.28	1.29	1.09	1.09
0.0005	1.17	1.21	1.25	1.28	1.30	1.32	1.34	1.36	1.11	1.11
0.001	1.19	1.24	1.28	1.31	1.34	1.36	1.38	1.40	1.12	1.12
0.005	1.25	1.32	1.37	1.42	1.45	1.48	1.51	1.54	1.16	1.16
0.010	1.30	1.38	1.44	1.49	1.53	1.57	1.60	1.63	1.19	1.19
0.020	1.36	1.46	1.53	1.59	1.64	1.69	1.73	1.76	1.23	1.23
0.025	1.39	1.49	1.57	1.63	1.69	1.74	1.78	1.82	1.24	1.24
0.050	1.49	1.62	1.73	1.81	1.88	1.95	2.00	2.05	1.33	1.33
0.075	1.58	1.74	1.87	1.97	2.05	2.13	2.20	2.26	1.41	1.41
0.100	1.67	1.86	2.00	2.12	2.22	2.31	2.38	2.45	1.49	1.49
0.20	2.05	2.34	2.57	2.76	2.92	3.06	3.18	3.30	1.82	1.82
0.30	2.52	2.95	3.29	3.58	3.82	4.03	4.22	4.38	2.24	2.24
0.40	3.18	3.82	4.33	4.75	5.11	5.42	5.70	5.95	2.83	2.86
0.50	4.20	5.17	5.95	6.59	7.14	7.63	8.06	8.44	3.74	3.88

Referencias

[1] ANGERS, C. "A Graphical Method to Evaluate Sample Sizes for the Multinomial Distribution". Technometrics, Vol.16, No. 3, pp. 469-471. 1974.

[2] "Large Sample Size for the Estimation of Multinomial Frequencies from Simulations Studies". Simulation: Oct, pp.175-178. 1984.

[3] COCHRAN, W.G. Técnicas de Muestreo. México D.F.: Continental, S.A. 1963.

[4] KISH, L. Muestreo de Encuestas. México D.F.: Trillas. 1972.

[5] MEDINA, F. "Tamaño Óptimo de Muestra en Encuestas de Propósitos Múltiples". En: CEPAL, Memoria del Taller Regional sobre Planificación de Encuestas en Hogares, Santiago de Chile. 1998.

[6] PÉREZ, C.). Técnicas de Muestreo Estadístico. Teoría, Práctica y Aplicaciones informáticas. México D. F.: Alfaomega. 2000.

[7] THOMPSON, K.T. Sample Size for Estimating Multinomial Proportions. The American Statistician, Vol. 41, No. 1, pp. 42-46. 1987.

[8] TORTORA, R.D. A Note on Sample Size Estimation for Multinomial Populations. The American Statistician Vol. 32, No. 3, pp. 100-103. 1978.