

**TAMAÑO ÓPTIMO DE MUESTRA EN ENCUESTAS DE PROPÓSITOS
MÚLTIPLES**

Fernando Medina

CEPAL

ÍNDICE

I. Introducción

II. Tamaño de muestra óptimo para distribuciones multinomiales: El caso univariado para diseños monoetápicos

1. El cálculo del tamaño de muestra: Aspectos prácticos

2. Ejemplo de cálculo

III. La determinación del tamaño de muestra en diseños polietápicos y multitemáticos

1. El planteamiento matemático

2. El criterio de optimalidad

3. Formulación específica: Muestreo aleatorio simple estratificado

4. Aplicación del método

5. Muestreo polietápico

6. Aplicación del método

IV. Conclusiones

Referencias bibliográficas

INTRODUCCIÓN

La determinación del tamaño de muestra en el diseño de una encuesta por muestreo probabilístico es una de las etapas más importantes, por lo que debe afrontarse con estricto apego a las consideraciones de carácter técnico de la teoría estadística, así como a los objetivos de la investigación y a los usos futuros de la información.

Resulta muy frecuente escuchar entre los especialistas en el tema que para calcular el número de unidades que formarán parte de la muestra hay que considerar la varianza de la variable de interés, así como la precisión con la que se desean obtener las estimaciones y la confianza requerida. Sin embargo, se pone muy poco énfasis en la importancia de tener presentes los objetivos de la encuesta, el tipo de variables e indicadores que se desean estimar, los dominios de estudio que se quieren analizar, así como el esquema de muestreo que se utilizará para seleccionar a las unidades de observación. Es cierto que el tamaño de muestra tiene un efecto sobre la varianza de las estimaciones (S^2/n); sin embargo, también se debe recordar que ésta, a su vez, depende en gran medida del esquema que se utilice para la selección de la muestra (Kish, 1965).

Es habitual que las encuestas que se realizan en la práctica sean de propósitos múltiples y por lo tanto se requiera la estimación de diversas estadísticas, las cuales, de manera individual, pueden conducir a tamaños de muestra y esquemas de selección diferentes. En este caso, la habilidad del especialista en muestreo conlleva a determinar el número de observaciones necesarias para cubrir los objetivos de un estudio multitemático, logrando generar un balance apropiado entre los costos de ejecución y la precisión deseada para los diferentes parámetros que se desean estimar. Sin embargo, se debe enfatizar que es muy difícil encontrar soluciones óptimas cuando las características que se desean investigar son demasiado dispares o las frecuencias observadas son muy bajas en los dominios de estudio de interés. Es obvio que las encuestas que se diseñan para estimar totales a niveles agregados y aquellas en donde se desean analizar subpoblaciones con características muy particulares requieren fracciones de muestreo muy diferentes.

En los textos que estudian la teoría del muestreo la determinación del tamaño de muestra para investigaciones multitemáticas se analiza, en primera instancia, como si se tratara de un problema univariado (Cochran, 1953 y Kish, 1965) sin considerar que prácticamente todas las encuestas son de propósitos múltiples. Así, en el caso particular de la determinación del tamaño de muestra en un estudio multipropósitos, cuando se desea estimar una proporción o porcentaje de una población -con función de densidad binomial- que posee ciertos atributos, se asume normalidad en la distribución de probabilidad del parámetro (p) y no se considera la corrección por población finita (cpf) (Cochran, 1953). Si al aplicar este procedimiento se ignora que otras características de la población también serán estimadas a partir de los datos recabados, no será posible determinar la precisión para las variables analizadas en forma simultánea y muy probablemente se

incrementará el error de muestreo.

Conforme a lo anterior, el objetivo de este trabajo es llamar la atención sobre la necesidad de asumir con solvencia técnica la etapa del cálculo del tamaño de muestra en las encuestas de propósitos múltiples y alertar sobre las limitaciones que se presentan en la confiabilidad de las estimaciones para dominios de estudio que no fueron considerados en los objetivos de la investigación. De manera particular, se señala la necesidad de evaluar la precisión de los indicadores que se obtienen y asumir con responsabilidad la formulación de hipótesis sobre el comportamiento de subpoblaciones muy específicas sobre las que no se puede garantizar que la inferencia estadística tenga validez.

II. TAMAÑO DE MUESTRA ÓPTIMO PARA DISTRIBUCIONES MULTINOMIALES: EL CASO UNIVARIADO PARA DISEÑOS MONOETÁPICOS

El cálculo del tamaño de muestra para estimar parámetros de proporciones con distribución multinomial es una situación cotidiana en el diseño de las encuestas que realizan las Oficinas Nacionales de Estadística.

De esta manera, el problema de determinar el número de observaciones necesarias para la estimación simultánea de proporciones multinomiales es equivalente a la construcción de intervalos de confianza simultáneos para variables con distribución multinomial, con la diferencia de que en el cálculo del tamaño de muestra los límites son fijados a priori por el investigador, a fin de controlar la probabilidad de que el intervalo contenga al verdadero valor del parámetro.

Cuando se realiza una encuesta para estimar parámetros de distribuciones multinomiales, el objetivo se debe centrar en calcular intervalos de confianza para cada una de las categorías de la variable. Sin embargo, en el trabajo cotidiano este hecho frecuentemente se pasa por alto, ya que en la práctica el procedimiento utilizado consiste en considerar a cada categoría - versus el resto-, como si se tratara de una variable binomial y se utiliza este hecho para determinar un conjunto de intervalos de confianza para cada una de las proporciones observadas en las celdas de manera independiente (Cochran, 1953).

Esta manera de actuar no es del todo apropiada ya que este procedimiento no permite el cálculo del coeficiente de confianza para el conjunto de intervalos, puesto que únicamente establece consideraciones sobre los valores de las proporciones analizadas de manera individual.

En Quesenberry y Hurts (QH, 1964) se propone un procedimiento para construir intervalos de confianza simultáneos para distribuciones multinomiales basados en una aproximación a la distribución chi-cuadrada (X^2) mediante el siguiente razonamiento.

Sean n_1, n_2, \dots, n_k las frecuencias observadas en una muestra de tamaño n con distribución de probabilidad multinomial y sean $\pi_1, \pi_2, \dots, \pi_k$ los parámetros asociados a la distribución. En este caso, n_i denota el número de observaciones clasificadas en la i -

ésima celda de la distribución, mientras que π_i representa la probabilidad de que alguna observación caiga en la i -ésima celda ($i=1, 2, \dots, k$), y es evidente que $\pi_i \geq 0$ y $\sum \pi_i = 1$.

Conforme a lo anterior, QH propusieron los siguientes intervalos de confianza simultáneos para los k parámetros de las $\pi_1, \pi_2, \dots, \pi_k$:

$$\pi_i^{i \in} \pi_i \pi_i^s \quad (i=1, 2, \dots, k) \quad (1)$$

donde:

$$\pi_i^{i \in} = \{ A + 2 n_i - \{ A [A + 4 n_i (n - n_i) / n] \}^{1/2} \} / [2 (n + A)]$$

representa el límite inferior y

$$\pi_i^s = \{ A + 2 n_i + \{ A [A + 4 n_i (n - n_i) / n] \}^{1/2} \} / [2 (n + A)]$$

permite el cálculo del límite superior del intervalo.

En las expresiones anteriores A representa el porcentaje superior ($\alpha * 100$) de una distribución chi-cuadrada (X^2) con $(k - 1)$ grados de libertad y n el tamaño total de la muestra ($A \sim X^2_{(k-1, \alpha)}$). Así, cuando n tiende a infinito ($n \rightarrow \infty$), la probabilidad de que los k intervalos de confianza contengan el verdadero valor del parámetro será menor a $(1 - \alpha)$.

Las expresiones para los intervalos de confianza definidas en (1) fueron recomendadas para el caso en que la probabilidad de que las aseveraciones acerca de los verdaderos valores π_i son correctas y deben ser mayores o iguales a $(1 - \alpha)$; es decir, cuando la probabilidad de que la afirmación sea incorrecta es menor o igual que α .

En un trabajo posterior, Goodman (1965) mejoró el método de cálculo de QH y propuso otra metodología que genera intervalos de confianza más pequeños, basados en la aproximación de la densidad binomial a la distribución normal, utilizando la desigualdad de Bonferroni y el teorema del límite central para determinar la probabilidad de que los intervalos sean correctos en forma simultánea.

Goodman demostró que los estimadores propuestos por QH se podrían obtener a partir de la solución de una ecuación cuadrática en π_i :

$$(p_i - \pi_i)^2 = A \pi_i (1 - \pi_i) / n \quad (i = 1, 2, \dots, k) \quad (2)$$

en donde $p_i = n_i / n$ y A tiene la misma interpretación que la sugerida por QH.

Cuando $k=2$ y ($n \rightarrow \infty$), la probabilidad de que los intervalos de confianza definidos por (1) y (2) contengan al verdadero valor del parámetro es $(1 - \alpha)$; y corresponden a los que se construyen para estimar los parámetros de una distribución binomial.

Cuando se trata de una variable multinomial, y en particular cuando $k > 2$, Goodman(*op. cit.*) demuestra que las expresiones (1) y (2) se pueden mejorar reemplazando A por B , en donde B corresponde al valor máximo del $(\alpha / 2) * 100$ percentil de una distribución chi-

cuadrada con un grado de libertad ($B \sim X^2_{(1, \alpha/2)}$).

Conforme a lo anterior, la probabilidad de que el intervalo de confianza obtenido sea incorrecto para π_i es igual a α/k ; mientras que la probabilidad de que al menos uno de los k intervalos de confianza sea incorrecto es igual a α .

Mediante este procedimiento se obtienen intervalos más pequeños que los propuestos por QH, y además según Goodman (*op. cit.*) este es el procedimiento apropiado para garantizar que la probabilidad de las afirmaciones sobre la precisión del intervalo sean correctas.

Las expresiones propuestas por QH para determinar intervalos de confianza en una muestra aleatoria de tamaño n proveniente de una distribución multinomial con parámetros p_1, p_2, \dots, p_n son:

$$p_i = p_i \pm Z_{(\alpha_i/2)} [p_i (1 - p_i) / n]^{1/2}; i = 1, 2, \dots, k. (3)$$

los cuales contendrán al verdadero valor del parámetro en forma simultánea con probabilidad $(1 - \sum \alpha_i)$, y donde p_i representa la proporción observada en la i -ésima celda y $Z_{(\alpha_i/2)}$ es el $(1 - \alpha_i/2) * 100$ percentil de una distribución normal estandarizada.

Posteriormente, Angers(1974) teniendo como referencia el trabajo de Goodman propuso un método gráfico para determinar tamaños de muestra para distribuciones multinomiales. Angers propuso que los puntos medios de los intervalos de confianza se pueden calcular utilizando la siguiente expresión:

$$d_i = Z_{(\alpha_i/2)} [p_i (1 - p_i) / n]^{1/2}; i = 1, 2, \dots, k (4)$$

A partir de (4) es posible calcular los valores de las d_i 's para determinados niveles de confianza α_i 's. Sin embargo, Angers (*op. cit.*) sugiere que es muy tedioso determinar el valor óptimo de n para las d_i 's, si el objetivo es que los k intervalos $p_i = p_i \pm d_i$ contengan al verdadero valor del parámetro en forma simultánea con un nivel de significancia $\alpha = \sum \alpha_i$. Así, la propuesta consiste en sugerir que el tamaño de muestra se determine de manera gráfica sustituyendo $\sum \alpha_i$ por $\sum \alpha_i/2$, por medio del siguiente procedimiento que se apoya en un razonamiento gráfico: elija de manera arbitraria un tamaño de muestra n y calcule los k cocientes por medio de la expresión:

$$n d_i^2 / p_i (1 - p_i); i = 1, 2, \dots, k \text{ (categorías de la variable de diseño) (5)}$$

que representan los valores de las abcisas, mientras que en el eje de las ordenadas se ubican los niveles de confianza ($0 \leq \alpha \leq .10$ y $0 \leq \alpha \leq .01$).

Posteriormente, se busca en las gráficas propuestas los valores obtenidos en el eje de las abcisas a fin de identificar los correspondientes niveles de confianza (α_i 's); y se compara la $\sum \alpha_i$ con el valor de α definido por el investigador, el criterio que se utiliza para decidir es que si la sumatoria ($\sum \alpha_i$) 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 al valor buscado. Cuando se logre ubicar el intervalo, el número final de observaciones se obtiene por medio de interpolación lineal simple.

1. El Cálculo del Tamaño de Muestra: Aspectos Prácticos

La relación de los procedimientos anteriores con la práctica cotidiana se puede ilustrar a partir del trabajo de Tortora (1978). Haciendo referencia a la manera en que Cochran (1953, pág. 50) aborda en su libro clásico la determinación del tamaño de muestra para estimar una proporción aplicando un esquema de selección aleatorio simple y para el caso univariado, Tortora retoma la propuesta de Goodman (*op. cit.*), y plantea el cálculo del tamaño de muestra de la manera siguiente.

Suponga que una antropóloga desea investigar el tipo de sangre de los habitantes de una isla y quiere seleccionar una muestra de tamaño n para estimar el porcentaje de los que tienen sangre de tipo O. Así, según el procedimiento descrito por Cochran (*op. cit.*), el experto en muestreo debiera considerar a la variable (tipo de sangre O) como si se tratara de eventos Bernoulli y el efecto del tamaño de muestra sobre el valor del estimador se debe medir en términos del error estándar o del intervalo de confianza. Sin embargo, de acuerdo al razonamiento de Tortora (*op. cit.*) suponga ahora que la antropóloga desea estudiar de manera simultánea la distribución de los distintos tipos de sangre de todos los habitantes de la isla.

Esto significa para el muestrista responder a la siguiente pregunta, ***qué tan grande debe ser la muestra para garantizar con suficiente precisión afirmaciones sobre la distribución de los porcentajes de los distintos tipos de sangre (A, B, O y AB) de los habitantes de la isla.***

La pregunta anterior debiera ser la que los muestristas de las Oficinas Nacionales de Estadística se plantearan y respondieran durante la etapa de elaboración de los diferentes diseños de muestra que involucran distribuciones multinomiales, ya que es muy común que en las encuestas se definan objetivos como los que se plantea la antropóloga hipotética del ejemplo de Cochran. Sin embargo, los hechos indican otra cosa, y claramente se enfrenta una inconsistencia entre el deber ser y la realidad, por lo que es posible que en algunas encuestas no se tengan observaciones suficientes para estimar con precisión parámetros de interés en subpoblaciones, con el agravante de que además en la divulgación, interpretación y utilización de los resultados existe poca preocupación de los usuarios por la precisión estadística de los datos y por supuesto las inferencias y decisiones de política que se plantean nunca hacen referencia al error de muestreo asociado a las estimaciones.

El procedimiento propuesto por Tortora (*op. cit.*) parte de dividir a una muestra de tamaño n en k categorías exhaustivas y mutuamente excluyentes. Así, sea π_i la proporción de la población ubicada en la i -ésima categoría ($\sum \pi_i = 1$) y n_i la frecuencia observada en la i -ésima categoría de una muestra aleatoria. Para una precisión α deseada, se quiere determinar el conjunto de intervalos I_i ($i = 1, 2, \dots, k$) tal que:

$$\Pr \left\{ \bigcap_{i=1}^k (\pi_i \in I_i) \right\} \geq (1 - \alpha) \quad (6)$$

Es decir, se desea que la probabilidad de que todo intervalo I_i contenga al verdadero valor del parámetro π_i sea menor a $(1 - \alpha)$.

A partir del trabajo de Goodman(*op. cit.*) se presentan expresiones que permiten aproximar el cálculo de intervalos de confianza cuando $n \rightarrow \infty$:

$$\pi_i^{i \pm} = \pi_i \pm [B \pi_i (1 - \pi_i) / n]^{1/2} \quad (7)$$

en donde:

$$\pi_i^i = \pi_i - [B \pi_i (1 - \pi_i) / n]^{1/2}$$

$$\pi_i^s = \pi_i + [B \pi_i (1 - \pi_i) / n]^{1/2}$$

representan el límite inferior y superior y B es el percentil superior $(\alpha / 2 * 100)$ de una distribución X^2 con 1 grado de libertad ($B \sim X^2_{(1-\alpha/2)}$).

El análisis de las ecuaciones dadas en (7) revelan que $[\pi_i (1 - \pi_i) / n]^{1/2}$ es la desviación estándar de la i -ésima celda de una población multinomial, y además se debe recordar que las funciones marginales de probabilidad corresponden a una densidad binomial. Así, si N es el tamaño total de la población y se incorpora el factor de corrección por población finita (cpf) y la varianza de cada π_i , se tiene una aproximación a los intervalos de confianza (Cochran, 1953).

$$\pi_i^i = \pi_i - [B (N - n) \pi_i (1 - \pi_i) / (N - 1) n]^{1/2}$$

$$\pi_i^s = \pi_i + [B (N - n) \pi_i (1 - \pi_i) / (N - 1) n]^{1/2} \quad (8)$$

Considerando que cuando $N \rightarrow \infty$ las expresiones anteriores convergen a (7).

Para la determinación del tamaño de muestra se requiere definir la precisión de cada parámetro de la distribución multinomial. De esta manera, suponga que se desea una precisión absoluta b_i para cada celda; entonces, a partir de (7) se tiene que:

$$\pi_i - b_i = \pi_i - [B \pi_i (1 - \pi_i) / n]^{1/2}$$

$$\pi_i + b_i = \pi_i + [B \pi_i (1 - \pi_i) / n]^{1/2} \quad (9)$$

Despejando el valor de b_i en las ecuaciones (9):

$$b_i = [B \pi_i (1 - \pi_i) / n]^{1/2} \quad (10)$$

y resolviendo para n se obtiene que el tamaño de muestra necesario para estimar cada celda con una precisión b_i es igual a:

$$n = \max_i \{ B \pi_i (1 - \pi_i) / b_i^2 \} \quad (11)$$

En caso que se desee incorporar la corrección por finitud (cpf) (11) se convierte en:

$$n = \max_i \{ B N \pi_i (1 - \pi_i) / [b_i^2 (N - 1) + B \pi_i (1 - \pi_i)] \} \quad (12)$$

Conforme a lo anterior, para decidir sobre qué tamaño de muestra resulta más apropiado se deben calcular los k pares (b_i, π_i) y seleccionar el valor de n que resulte mayor dejando el valor de B fijo en cada caso.

Es importante observar que en (11) y (12) el tamaño de muestra es una función de π_i y de b_i , por lo que n decrece cuando $\pi_i \rightarrow 1/2$ y también lo hace en el caso en que $b_i \rightarrow 0$. En la situación en que $b_i = b \forall i$, sólo se requiere hacer un cálculo y se sugiere que este sea con la π_i más cercana a $1/2$.

Es frecuente que en la práctica se desconozca un valor aproximado de la proporción π_i que se desea estimar; por lo que se puede suponer -como frecuentemente se hace- que $\pi_i = 1/2$ y $b_i = b \forall i = 1, \dots, k$, por lo que $n = B / 4 b^2$.

2. Ejemplo de Cálculo

A fin de ejemplificar el procedimiento propuesto por Tortora, continuemos desarrollando el ejemplo hipotético de la antropóloga que desea estimar las proporciones de habitantes con tipos de sangre A, O, B y AB. Para este fin, supóngase que de un estudio previo se sabe que la distribución de los tipos de sangre en una población similar fue la que se presenta en el cuadro 1.

Cuadro 1
DISTRIBUCIÓN DEL TIPO DE SANGRE

TIPO DE SANGRE	PORCENTAJE
A	27
B	19
O	43
AB	11

Además, suponga que se desea obtener una precisión absoluta de $\pm 5\%$ para cada proporción y un coeficiente de confianza de .95 ($\alpha = 5\%$). Según la notación utilizada por Tortora, en este caso $b_i = .05$ ($i = 1, 2, 3$ y 4) ya que se requiere la misma precisión para cada celda y $\alpha = .05$ ($(1 - \alpha) = .95$), y como supuesto adicional se asume que la población es suficientemente grande por lo que se ignora la cpf.

Debido a que $b_i = b \forall i$, entonces sólo debemos efectuar un cálculo utilizando la proporción que se encuentre más cerca de .5, y que en este caso corresponde a la

proporción de la población con tipo de sangre O. Así, utilizando la expresión (11) y sustituyendo de forma conveniente:

$$n = B \pi_i (1 - \pi_i) / b_i^2 = 6.5381 * 0.43 (1 - 0.43) / (.05)^2 = 641$$

se obtiene que el tamaño de muestra requerido es de 641 personas. En este caso, $k=4$ y B es igual a 6.5381, que corresponde al valor en tablas de una distribución $X^2_{(1-\alpha/k=0.0125)}$ con un grado de libertad para un valor de $\alpha / k = 0.0125$

(1.25%) ; ya que $B \sim X^2_{(1-\alpha/k)}$.

Por otra parte, si el problema se resolviera de la manera tradicional se tendrían que calcular tres tamaños de muestra diferentes -uno para cada valor de p - suponiendo en cada caso una distribución binomial (normal) del parámetro de interés.

Así, considere la popular expresión para el cálculo del tamaño de muestra para una proporción binomial:

$$n' = t^2 p q / d^2 \quad (13)$$

en donde t representa la abcisa de la curva normal para una confianza α determinada, mientras que d es el error máximo absoluto con el que se desean obtener las estimaciones. De esta manera, con $\alpha = .05$ los tamaños de muestra calculados para cada una de las proporciones requeridas son:

$$n_A = 303 ; n_B = 236 ; n_O = 377 \text{ y } n_{AB} = 150$$

Según el criterio definido por Cochran se debiera escoger el valor máximo y seleccionar 377 personas; sin embargo, observe que este valor representa tan sólo el 58.8% del total de observaciones calculadas mediante el método propuesto por Tortora. Es decir, se estaría seleccionado una muestra 41.2% menor de lo que realmente se necesita. Ante esta situación, es evidente que se incrementa el error de muestreo cuando se intentan desagregar las observaciones para analizar características de interés en dominios de estudio específicos con baja frecuencia de aparición.

La comparación de ambas fórmulas se logra por medio del cociente de las expresiones (11) y (12) el cual indica que:

$$n / n' = B / t^2 \quad (14)$$

Observe que la relación varía de acuerdo al valor que asume $B \sim X^2_{(1-\alpha/k)}$, el cual a su vez depende de la precisión requerida (α) y del número de categorías en que se distribuye la variable de interés (k). De esta manera, es posible construir una tabla en donde se establezca la relación que existe entre el tamaño de muestra de una distribución binomial y aquel que se obtiene por medio de intervalos de confianza simultáneos para distribuciones multinomiales.

Cuadro 2

RELACIÓN ENTRE EL TAMAÑO DE MUESTRA BINOMIAL Y DE INTERVALOS DE CONFIANZA SIMULTÁNEOS MULTINOMIALES¹

CATEGORÍAS CONFIANZA α	3^2	4	5	6	7	8	9	10
.010	1.30	1.38	1.44	1.48	1.54	1.56	1.60	1.63
.025	1.39	1.49	1.57	1.63	1.69	1.74	1.78	1.82
.050	1.49	1.62	1.73	1.81	1.89	1.94	2.00	2.05
.075	1.58	1.74	1.87	1.97	2.05	2.13	2.20	2.26
.100	1.68	1.86	2.00	2.12	2.22	2.31	2.38	2.45

¹ En las celdas se reportan los valores del cociente $n / n' = B / t^2$ y no se considera la corrección por población finita (cpf), además de que se asume la misma precisión absoluta para cada una de las categorías.

² Cuando $k=2$, se trata de una distribución binomial y el tamaño de muestra se determina de la manera tradicional propuesta por Cochran.

La interpretación del cuadro 2 se debe hacer de la manera siguiente. Supóngase que se desean estimar las proporciones de una cierta variable con distribución de probabilidad multinomial formada por 3 categorías y se quiere garantizar una confianza de $(1 - \alpha) = 95\%$ y la misma precisión para todas las celdas. En este caso, la fórmula propuesta por Tortora indicaría que se tendría que seleccionar una muestra 49% mayor a la que se determinaría si se aplicara la expresión tradicional que se presenta en Cochran(1953). Esto significa que (13) subestimaría el verdadero tamaño de la muestra necesario para obtener la precisión deseada en cada una de las tres categorías en que se clasifica la variable de interés a una confianza del 95%.

El método de Tortora representa la manera correcta de proceder para determinar el tamaño de muestra de una proporción multinomial bajo un esquema de selección aleatorio simple y cuando se utiliza sólo una variable de diseño; sin embargo, en Angers (1979) se mejora esta propuesta ya que el autor considera que en ocasiones los resultados generados por la expresión (11) son muy conservadores y estiman tamaños de muestra innecesariamente grandes.

Angers (*op. cit.*), basa su procedimiento usando para el cálculo un valor muy cercano a .5

para cada parámetro mediante la siguiente fórmula:

$$n = \min \alpha_i, \max_i \{ B_i \pi_i (1 - \pi_i) / b_i^2; i = 1, \dots, k \} \quad (15)$$

tal que $\alpha_i \leq a_i \leq \alpha$ y $\sum_{i=1}^k \alpha_i \leq \alpha$. En esta propuesta, $B \sim X^2_{(1, a/k)}$ tiene la misma interpretación que en el método de Tortora y las a_i 's = α / k y el valor de α son propuestos por el investigador. Así, la probabilidad de que el i -ésimo intervalo de confianza contenga al verdadero valor del parámetro t_i es al menos $(1 - \alpha_i)$, mientras que $(1 - \alpha)$ representa la probabilidad de que los k intervalos sean correctos de manera simultánea.

Según los resultados de las aplicaciones que presenta Angers en su investigación, esta forma de calcular el tamaño de muestra es más robusta y puede llegar a significar ahorros de casi un 12% en el número de observaciones a seleccionar en relación al método de Tortora.

Thompson (1987) hace una revisión de las investigaciones realizadas y concluye que el método propuesto por Angers(1984) supera las propuestas anteriores pero resulta muy tedioso en su aplicación. En este sentido, el autor propone una manera de determinar el "peor de los casos" (worst case) para un vector de parámetros multinomiales cuando se desean obtener intervalos de confianza simultáneos para cada uno de los componentes del vector p .

Thompson 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 $(1 - \alpha)$; esto es,

$$\Pr \{ \bigcap_{i=1}^k | p_i - \pi_i | \leq d_i \} \geq 1 - \alpha \quad (16)$$

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

$$a_i = \Pr \{ \frac{1}{2} Z_i \leq d_i \sqrt{n} / \sqrt{p_i (1 - p_i)} \} = 2(1 - \Phi(Z_i)) \quad (17)$$

en donde Z_i es la variable normal estandarizada, Φ la función acumulativa de probabilidad y $Z_i = d_i \sqrt{n} / \sqrt{p_i (1 - p_i)}$.

A fin de obviar la corrección por finitud y lograr la normalidad del estimador, se asume que la población es lo suficientemente grande y que el muestreo se hace sin reemplazo. En la expresión (16) y siguiendo a Angers (1974), se define las d_i como:

$$d_i = Z_{(\alpha_i/2)} [p_i (1 - p_i) / n]^{1/2} \quad (18)$$

en donde $Z_{(\alpha_i/2)}$ es el percentil $(1 - \alpha_i/2) * 100$ de una distribución normal estandarizada.

Como resultado de esta investigación, Thompson demuestra que para un determinado nivel de significancia α_i de las d_i 's (excepto cuando $\sum_i^k \alpha_i = \alpha$), el vector de parámetros de una distribución multinomial para determinar "el peor de los casos" se maximiza para \sum

$\alpha_i = \alpha$, cuando $p_i = 1/m$ para un número determinado de categorías y $p_i = 0$ para el resto; es decir,

$$p \pm Z_{(\alpha/2)} \left[\frac{1}{m} \left(1 - \frac{1}{m} \right) / n \right]^{1/2}, i = 1, \dots, k \quad (19)$$

en donde los valores de m varían conforme a k y a .

Conforme a lo anterior, y suponiendo que $d_i = d$ $i=1, \dots, k$ y sin tener ninguna información acerca de la forma en que se distribuye la variable en las distintas categorías, Thompson (*op. cit.*) propone la siguiente expresión para el cálculo del tamaño de muestra:

$$n = \max_m Z^2 \left(\frac{1}{m} \right) \left(1 - \frac{1}{m} \right) / d^2 \quad (20)$$

En donde $Z = j(1 - a/2m)$ es el límite superior del $(a/2m)$ percentil de una distribución normal y $m \hat{=} Z^2$.

Thompson sugiere que el tamaño de muestra se puede determinar a partir del cuadro 3 dividiendo $d^2 n / d^2$, cuando el investigador especifica un nivel de confianza a y una distancia (error de estimación) d .

Para ilustrar el uso del cuadro 3 suponga que se desea determinar el tamaño de muestra para realizar una encuesta que pretende estudiar el comportamiento de las mujeres en edad fértil por grupos de edad en función al número de hijos nacidos vivos. Así, si se desea una confianza de 95% para todas las categorías, entonces se fija $a = 5\%$ y en el cuadro 3 se observa que se deberían seleccionar un total de 510 mujeres ($d^2 n = 1.27359$), mientras que en la última columna se estaría indicando el número de parámetros distintos de cero. Es importante notar que no se requiere conocer a priori el número de grupos de edades que se formarán, lo cual significa que bajo esta manera de proceder no es necesario saber el número de categorías (k) en que se divide la población.

Cuadro 3

Tamaño de Muestra para Estimar en forma Simultánea los parámetros de una distribución Multinomial con una distancia "d" del verdadero valor con un nivel a de significancia

a	$d^2 n$	ncon $d=.05$	m
.50	.44129	177	4
.40	.50729	203	4
.30	.60123	241	3

.20	.74739	299	3
.10	1.00635	403	3
.05	1.27359	510	3
.025	1.55963	624	2
.02	1.65872	664	2
.01	1.96986	788	2
.005	2.28514	915	2
.001	3.02892	1212	2
.0005	3.33530	1342	2
.0001	4.11209	1645	2

Reproducida de Thompson (*op. cit.*)

A partir del análisis realizado se puede concluir que cuando se desean estimar parámetros de distribuciones multinomiales y garantizar la precisión simultánea de todos ellos, la expresión propuesta por Cochran no resulta apropiada ya que subestima de manera importante el número de observaciones necesarias para garantizar la precisión deseada en todas las categorías en las que se distribuye la variable de interés.

Asimismo, es importante enfatizar que en la práctica es habitual que se deseen estimar proporciones multinomiales por lo que se aconseja aplicar procedimientos robustos que están disponibles y que superan a los que se utilizan de manera tradicional para determinar el tamaño de muestra ($n=t^2pq/d^2$).

III. LA DETERMINACIÓN DEL TAMAÑO DE MUESTRA EN DISEÑOS POLIETÁPICOS Y MULTITEMÁTICOS

Hasta ahora, en los desarrollos presentados además de los supuestos en los que se basan los métodos se asume que la selección de las unidades de observación se realiza en una sola etapa y se trabaja como si sólo se deseara estimar una característica de la población. Es decir, se supone un esquema de muestreo monoetápico y aleatorio simple (MAS), en donde únicamente existe una variable de interés (proporción, media o total). Sin embargo, en la práctica no es común que se presenten estas situaciones ya que en la mayoría de los diseños de muestra al menos se requieren dos etapas para seleccionar a las

unidades de observación, y prácticamente todas las encuestas se pueden considerar de propósitos múltiples.

Lo anterior significa que a los problemas derivados de la falta de aplicación de un procedimiento apropiado para la determinación del tamaño de muestra se agrega el hecho de que el número de observaciones a seleccionar se debe incrementar debido a la necesidad de formar estratos y conglomerados para la selección de las unidades primarias de muestreo (UPM's), lo cual aumenta la varianza de los estimadores.

En la práctica, el problema de seleccionar la muestra en varias etapas y formar estratos y conglomerados para facilitar la identificación de las unidades de observación, se resuelve corrigiendo el tamaño de muestra obtenido mediante muestreo aleatorio simple por un factor denominado efecto de diseño (**efd**) que relaciona el coeficiente de correlación intraconglomerados con el tamaño promedio de las unidades de segunda etapa.

En Cochran (*op. cit.*), se señala que es habitual que en la planeación de las encuestas las organizaciones ejecutoras estén interesadas en estudiar múltiples características de la población, por lo que propone que para determinar el tamaño de muestra se fije el error máximo aceptable para cada una de las características que se consideren más importantes y se calcule de manera independiente el número de observaciones requeridas para cada caso.

Posteriormente, se sugiere comparar los diversos valores obtenidos y decidir al respecto: si éstos son similares y el tamaño de muestra más grande está acorde con el presupuesto destinado para la realización de la encuesta, se opta por este número de observaciones el cual garantizará suficiente precisión para todas las variables de interés. Sin embargo, la práctica cotidiana revela que esta situación es poco frecuente y en muchas ocasiones los valores son tan discordantes que es necesario renunciar a investigar algunas características de la población.

Es difícil que los diseñadores de encuestas con poca experiencia -y sobre todo los usuarios de la información- comprendan esta situación, y por lo tanto es frecuente que se decida investigar características de la población en donde el tamaño de muestra utilizado no está en condiciones de estimar con la precisión adecuada, sin que exista preocupación por evaluar la magnitud del error de estimación.

Asimismo, cuando se desean analizar subclases con características poco frecuentes es difícil estimar el tamaño de muestra ya que la distribución de la variable se conoce a posteriori y es muy probable que se incremente el error de muestreo de manera considerable (Yates, 1953). Esta situación es cotidiana en las encuestas de hogares (empleo, niveles de vida y presupuestos familiares) sin que los usuarios de la información intuyan que la calidad de los resultados que analizan y las recomendaciones de política que formulan pueden carecer de validez estadística (CEPAL, 1998).

De esta manera, los problemas de subestimación del tamaño de muestra se agravan con la consiguiente pérdida de precisión e incremento en el error de muestreo y sesgo de los estimadores.

En los textos sobre la teoría del muestreo se trata ampliamente el tema de la

determinación del tamaño de muestra para diseños que requieren obtener el número óptimo de selecciones de primera y segunda etapa para un presupuesto determinado o una varianza deseada, y además el interés del investigador se centra en estudiar una sola variable (Cochran, 1953, Snedecor y Cochran, 1980). Sin embargo, la mayoría de las encuestas requieren la estimación simultánea de varias variables con una precisión adecuada y en este caso los procedimientos para calcular el tamaño de muestra óptimo no están bien definidos.

1. El Planteamiento Matemático

Ya se señaló que la manera tradicional de proceder consiste en estimar el tamaño óptimo para cada variable -utilizando la expresión propuesta por Cochran y posteriormente decidirse por un valor que de alguna manera signifique una solución que satisfaga los requerimientos en forma conjunta (denominada solución de compromiso); sin embargo, esta no es la forma más adecuada de resolver el problema ya que frecuentemente no se logra obtener la solución óptima. De hecho, cuando se trata de estimar de manera simultánea un determinado número de características ($p \geq 2$) no existe un criterio único para la determinación del tamaño de muestra óptimo (Kokan, 1963), por lo que actuar de la manera tradicional no garantiza la confiabilidad requerida en las diversas variables de interés.

No existen muchas investigaciones que traten este tema y es preciso señalar que la solución matemática para determinar el número óptimo de selecciones de primera y segunda etapa cuando el interés se centra en la estimación simultánea de varias variables, significa minimizar una función de costos no lineal sujeta a restricciones no lineales y varianza constante lo cual puede resultar bastante complejo.

Si el análisis se centra en una variable el procedimiento para determinar el tamaño óptimo de muestra para diseños de una etapa está bien documentado en la literatura. Es decir, cuando se desea determinar el número óptimo de selecciones que minimice el costo de la encuesta para una varianza deseada ($V = a_0 + \sum_{i=1}^k a_i/m_i$) o minimizar la varianza del estimador para un presupuesto determinado ($C = c_0 + \sum_{i=1}^k c_i m_i$); en donde mas m_i 's son función del número de selecciones y $k \in \mathbb{Z}^+$ depende del procedimiento de muestreo seleccionado; c_i es el costo por muestrear la unidad m_i y a_i no depende de m_i pero si de su variabilidad.

Así, Kokan (*op. cit.*) demuestra que para un costo C fijo el valor mínimo de la varianza se obtiene por medio de:

$$V_{\min} = a_0 + \left(\sum_{i=1}^k \frac{a_i c_i}{C - c_0} \right)^2 / (C - c_0) \quad (21)$$

Por otra parte, cuando la varianza V es fija el valor mínimo de C se obtiene por medio de:

$$C_{\min} = c_0 + \left(\sum_{i=1}^k \frac{a_i c_i}{V - a_0} \right)^2 / (V - a_0) \quad (22)$$

y los tamaños de muestra óptimos son:

$$m_i = \frac{a_i c_i (C - c_0)}{\sum_{i=1}^k a_i c_i} ; (i=1, \dots, k) \text{ (para } C \text{ fijo)} \quad (23)$$

$$m_i = \frac{a_i}{c_i} \left(\frac{a_i}{c_i} \right)^{k_0} / (V - a_0) ; (i=1, \dots, k) \text{ (para } V \text{ deseada) (24)}$$

En la literatura no existe una solución tan simple cuando se trata de una encuesta que estudia varias características de manera simultánea, ya que las a_i 's cambian y es probable que los costos c_i también dependiendo de la variable de estudio, y por lo tanto no es posible encontrar un método único que minimice la varianza de todas las variables en forma simultánea para un presupuesto determinado.

Existen algunas propuestas para solucionar esta situación (Chakravarthy, 1955 y Ghosh, 1958) que han planteadas por Kokan (*op. cit.*) como un problema de optimización de una función sujeta a restricciones mediante técnicas de programación no lineal de la manera siguiente.

2. El Criterio de Optimalidad

Suponga que se dispone de un determinado presupuesto C para la realización de una encuesta en la cual se desean investigar p características de una población de interés. Es bien sabido que no se puede obtener un valor exacto en las estimaciones por lo que se está dispuesto a aceptar un cierto margen de error el cual dependerá de la característica de estudio, y para el cual se puede fijar una cota superior a la varianza (V) con algún nivel de confianza α .

Así, un vector $\mathbf{m} = (m_1, m_2, \dots, m_k)$ se dice que es óptimo si minimiza el costo de la investigación sujeto a la condición de que la varianza V_j de la j -ésima característica de interés no sea mayor a una cantidad v_j prefijada, para un nivel de confianza $1 - \alpha_j$ ($0 < \alpha_j < 1$), $j=1, \dots, p$, y en donde p representa el número de variables a estudiar.

Siguiendo a Kokan (*op. cit.*) la solución general se puede plantear de la manera siguiente. Considere una población de tamaño N (U_1, U_2, \dots, U_N) en donde se desean medir p características variables. Se determina un esquema de selección y se extrae una muestra de tamaño n y se fija la varianza V_j para la j -ésima variable. De esta manera, el problema de asignación óptima de la muestra consiste en minimizar una función de costos C que depende del vector \mathbf{m} sujeta a las restricciones de varianza (V_j), lo cual se plantea de la manera siguiente:

$$\text{Mín } C = f(\mathbf{m}) \text{ (25)}$$

$$\text{s.a. } V_j \leq v_j ; (j=1, \dots, p) \text{ (26)}$$

$$0 \leq m_i \leq M_i ; (i=1, \dots, k) \text{ (27)}$$

Si existe un vector \mathbf{m} que satisface las $(p+2k)$ restricciones impuestas por (26) y (27) se dice que existe una solución factible. Además, si esa solución minimiza la función objetivo (25) se obtiene un mínimo local óptimo.

3. Formulación Específica: Muestreo Aleatorio Simple Estratificado

Suponga que se forman L estratos y que n_h unidades son seleccionadas sin reemplazo del estrato h , y sin pérdida de generalidad suponga que se está interesado en estudiar la media de la población \bar{Y}_j . Así, un estimador insesgado de la media poblacional \bar{Y}_j es \bar{y}_{jst}

el cual tiene una varianza muestral igual a :

$$V_j = \text{var}(\hat{y}_{j\text{est}}) = \hat{a}_h^L a_{hj} x_h \quad (28)$$

$$\text{en donde } a_{hj} = W_h^2 S_{hj}^2 \quad (h=1, \dots, L; j=1, \dots, p) \quad (29)$$

$$\text{y } x_h = (1/n_h) - (1/N_h) ; \quad (h=1, \dots, L) \quad (30)$$

Si c_h es el costo de evaluar las p características de interés en una unidad cualquiera del h -ésimo estrato, entonces el costo total de la encuesta se determina por medio de la siguiente expresión :

$$C = c_0 + \hat{a}_h^L c_h n_h \quad (31)$$

en donde c_0 es un costo fijo. Si se aplican (29) y (30) en (28), la función de costo se transforma en:

$$C = c_0 + \hat{a}_h^L N_h c_h / (1 + N_h x_h) \quad (32)$$

Por consideraciones de carácter práctico en cada uno de los estratos el tamaño de muestra debe ser mayor o igual que uno (si fuera igual a cero no tendría sentido formar el estrato) y menor al tamaño del estrato, lo cual implica que:

$$0 \leq x_h \leq 1 - (1/N_h) \quad (33)$$

La situación ahora consiste en determinar los valores n_1, \dots, n_L con objeto de minimizar (31) sujeto a $V_j \leq v_j$ y a (33), para un costo c_0 fijo que puede no ser considerado en la solución del problema. **Minimizar C es lo mismo que maximizar $-C$** por lo que se formula el siguiente planteamiento:

$$\text{Max } f = - \hat{a}_h^L N_h c_h / (1 + N_h x_h) \quad (34)$$

$$\text{s.a. } \hat{a}_h^L a_{hj} x_h \leq v_j \quad (j=1, \dots, p) \quad (35)$$

$$0 \leq x_h \leq 1 - 1/N_h \quad (h=1, \dots, L) \quad (36)$$

Observe que (34) es una función no lineal y convexa, mientras que (35) y (36) son lineales en las x_h 's, y los valores V_j deben ser fijados por el investigador.

4. Aplicación del Método

Suponga que se desea realizar una encuesta para estimar el volumen de la producción (X_1) y del personal ocupado (X_2) de la industria manufacturera en un país determinado. Se dispone de un directorio de establecimientos elaborado a partir del Censo Económico más reciente y se cuenta con información sobre el tamaño del establecimiento lo cual, a partir de la aplicación de un procedimiento multivariado apropiado, permite la formación de tres estratos como se muestra a continuación:

**Clasificación de los Establecimientos de la Industria
Manufacturera por Tamaño**

Tamaño del Establec.	h	No. Estab. N_h	Varianza de la producción (S^2_{h1}) X_1	Varianza del no. de Empleados (S^2_{h2}) X_2	W_h	W_h^2
Chicos	1	800	4,000	12	.2857	.0816
Medianos	2	1,500	46,500	35	.5357	.2870
Grandes	3	500	500,000	250	.1786	.0319
Total		2,300			1.0000	

Las varianzas para los estimadores del total X_1 y X_2 se obtienen por medio de:

$$V_1 = \text{var}(x_1) = \sum_{h=1}^3 N_h^2 S_{h1}^2 x_h \quad \text{y} \quad V_2 = \text{var}(x_2) = \sum_{h=1}^3 N_h^2 S_{h2}^2 x_h$$

$$x_1 = 1/n_1 - (1/N_1) = 1/n_1 - (1/800) = 1/n_1 - .00125$$

$$x_2 = 1/n_2 - (1/N_2) = 1/n_2 - (1/1,500) = 1/n_2 - .000666$$

$$x_3 = 1/n_3 - (1/N_3) = 1/n_3 - (1/500) = 1/n_3 - .0020$$

Se desea que la diferencia entre el valor real y el estimador de la producción total (x_1) no sea mayor a 500,000, y que la diferencia entre el total de empleados estimados (x_2) y el total real se ubique alrededor de las 1,200 unidades. Además, se quiere que el **error de estimación** no sea mayor al **5%** con un nivel de probabilidad $p=0.95$. De esta forma, suponiendo normalidad de la variable se tiene que;

$$\Pr \left\{ \frac{1}{2} y_j - Y_{j\pm\epsilon} \leq .05 Y_j \right\} = \Pr \left\{ \frac{1}{2} y_j - Y_{j\pm\epsilon} \leq S.E(y_j) * 1.96 \right\} = .95 \quad j = 1, 2.$$

Además, para los fines del estudio se requiere que las varianzas de las estimaciones sean: $V_{1\epsilon} 150'000,000$ ($D.S(y_1)=12,247.45$) y $V_{2\epsilon} 80,000$ ($D.E(y_2)=282.84$), y sin incorporar restricciones adicionales se asume que no existen diferencias en los costos de entrevista en los diferentes estratos. De esta forma, se tiene que $C = c_0 + c \sum_{h=1}^3 n_h$ ($c_1=c_2=c_3=c$) y como los costos no afectan la solución del problema se define la siguiente función a maximizar:

$$\text{Max } f = \sum_{h=1}^3 n_h = - [(800/1+800x_1) + (1,500/1+1,500x_2) + (500/1+500x_3)]$$

$$\text{s.a } 2.56 X_1 + 104.625 X_2 + 125 X_3 \leq .15$$

$$7.68 X_1 + 78.75 X_2 + 6.25 X_3 \leq .08$$

$$0 \leq X_1 \leq .99875; 0 \leq X_2 \leq .999334; 0 \leq X_3 \leq .998000$$

Los resultados obtenidos mediante la aplicación del algoritmo de optimización indican que $x_1=.003153$, $x_2=.000708$ y $x_3=0$, por lo que se deben seleccionar un total de $n=1,454$ (63.2% del total) empresas distribuidas como se muestra a continuación: $n_1=227$, $n_2=727$ y $n_3=500$. Obsérvese que la solución obtenida señala que hay que censar el estrato de las empresas mayores que es donde se observa mayor variabilidad lo cual encarece el costo de la encuesta.

En una investigación posterior realizada por Kokan y Khan (1967) el problema presentado se replantea de la manera siguiente: si se asume que $x_i= 1/n_i$, entonces las ecuaciones (33), (34) y (35) se pueden reescribir como se muestra a continuación:

$$\text{Max } f = - \sum_{h=1}^L N_h c_h / (1+N_h x_h) \quad (37)$$

$$\text{s.a. } \sum_{h=1}^L a_{hj} x_h \leq v_j + \sum_{h=1}^L a_{hj} / N_h; (j=1,\dots,p) \quad (38)$$

$$1/N_i \leq x_h \leq 1; (h=1, \dots, L) \quad (39)$$

Conforme a lo anterior, se reestimaron los tamaños de muestra del ejemplo que estamos analizando a partir de la optimización del siguiente sistema de ecuaciones:

$$\text{Max } f = \sum_{h=1}^3 n_h = - [(800/1+800x_1) + (1,500/1+1,500x_2) + (500/1+500x_3)]$$

$$\text{s.a } 2.56 X_1 + 104.625 X_2 + 125 X_3 \leq .15+.0032+.06975+.25$$

$$7.68 X_1 + 78.75 X_2 + 6.25 X_3 \leq .08+.0096+.0525+.125$$

$$.00125 \leq X_1 \leq 1; .0000666 \leq X_2 \leq 1 \text{ y } .002 \leq X_3 \leq 1$$

A partir de la aplicación del algoritmo de optimización no lineal se tiene que $x_1=0.014385$, $x_2=0.001779$ y $x_3=0.002$, por lo que los tamaños de muestra estimados para cada estrato son $n_1=64$, $n_2=409$ y $n_3=250$. De esta forma, la solución encontrada indica que se deben seleccionar un total de $n=723$ empresas (25.8% del total) de la industria manufacturera, con objeto de obtener estimaciones del volumen de la producción y del total de personal ocupado con el nivel de confianza requerido.

Es importante señalar que existen diferencias importantes respecto a la manera en que se procede comúnmente. Así, es habitual que primero se decida el número total de observaciones que se van a seleccionar (n) y posteriormente se distribuya la muestra con algún criterio de asignación entre los diferentes estratos. En el procedimiento aplicado, el total de unidades que formarán parte de la investigación se determina a partir de agregar los tamaños de muestra calculados para cada uno de los estratos considerando la variabilidad de la variable de interés y las diferencias de tamaño.

5. Muestreo Polietápico

Suponga que se decide un esquema de selección en dos etapas en donde las unidades primarias de muestreo (UPM's) son de igual tamaño y que se está interesado en estimar p características de las unidades de observación. Se asume, sin pérdida de generalidad, el interés por estimar la media \bar{y}_j cuya varianza muestral tiene la siguiente expresión:

$$V_j = \text{var}(\bar{y}_j) = (1/n - 1/N) S_{uj}^2 + (1/nm - 1/NM) S_{wj}^2 \quad (40)$$

Se define $x_1 = (1/n - 1/N)$ y $x_2 = (1/nm - 1/NM)$, de tal manera que $n = N/(1+Nx_1)$

y $nm = NM/(1+NMx_2)$. Como $1 \leq n \leq N$ y $1 \leq m \leq M$, se tiene que $0 \leq x_1 \leq 1-1/N$ y

$0 \leq x_2 \leq 1-1/NM$. Así, si definimos a FC como la función de costos:

$$FC = c_0 + c_1 n + c_2 nm + c_3 n^{1/2} \quad (41)$$

En donde c_0 es un costo fijo y c_1, c_2 representan los costos de entrevista en las unidades de primera y segunda etapa respectivamente, mientras que c_3 es el costo de traslado entre las UPM's y USM's.

Debido a que c_0 es una cantidad constante y fijando el límite superior de la varianza esperada $V_j = v_j$, se plantea el problema de maximizar:

$$\text{Max } f = - [(c_1 N / (1 + N x_1)) + (c_2 N M / (1 + N M x_2)) + (c_3 N^{1/2} / (1 + N x_1))] \quad (42)$$

$$\text{s.a. } S_{u1}^2 x_1 + S_{w1}^2 x_2 \leq v_1$$

$$\dots \dots \dots (43)$$

$$S_{up}^2 x_1 + S_{wp}^2 x_2 \leq v_p$$

$$0 \leq x_1 \leq 1-1/N, 0 \leq x_2 \leq 1-1/NM \quad (44)$$

6. Aplicación del Método

Suponga que se realizó una encuesta para estudiar el tema de la desocupación en donde se seleccionaron 25 unidades de primera etapa (n) y 10 de segunda etapa (m) con igual probabilidad, y se estimaron las siguientes varianzas para las variables x_1 y x_2 que representan la tasa de desocupación abierta y la de subempleados en una determinada ciudad.

$$S_{i1}^2 = .0087, S_{i2}^2 = .0432, S_{e1}^2 = .0492 \text{ y } S_{e2}^2 = .0850$$

en donde los subíndices i y e indican que se trata de la varianza intraconglomerados y entre conglomerados respectivamente. En este caso, $n=25$ y $m=10$ por lo que $nm=250$, mientras que el costo de entrevista en las UPM's es $c_1=\$7.8$ y en las USM's se estimó en $c_2=\$3$, mientras que $c_3=0$.

Suponga ahora que la Oficina Nacional de Estadística (ONE) desea realizar una encuesta y para la elaboración del nuevo diseño de muestra requiere **determinar los valores**

óptimos de n y m de tal forma que se minimice el costo total de la investigación, ya que se cuenta con un presupuesto limitado y se desea utilizar como insumos las varianzas estimadas de la investigación anterior. Observe que no se conocen los valores de N y M los cuales se pueden suponer como infinitos lo cual no impone ninguna restricción adicional a la formulación y solución del problema.

Conforme al procedimiento descrito en la sección anterior se plantea el siguiente problema de optimización.

$$\text{Max } \mathbf{f} = - [7.8/x_1 + 3/x_2]$$

$$\text{s.a. } S_{u1}^2 + S_{w1}^2 \mathbf{f} v_1$$

$$S_{u2}^2 + S_{w2}^2 \mathbf{f} v_2 \quad (45)$$

$$0 < x_1 = 1/n \mathbf{f} 1$$

$$0 < x_2 = 1/nm \mathbf{f} 1$$

Para determinar el límite superior de V_j ($j=1,2$) se utiliza la siguiente expresión propuesta por Kokan(1963):

$$\text{Pr } \{ V \mathbf{f} (n-1)s_e^2 / (nmC^2_0) \} = 1 - a \quad (46)$$

En este caso, como se desea un $a = .05$ entonces $C^2_{0(n-1, a)} = C^2_{0(24, .05)} = 13.85$, por lo que los valores de v_1 y v_2 se obtienen como se indica a continuación:

$$v_1 = (24)(.0492)/(250)(13.85) = .000341$$

$$v_2 = (24)(.0850)/(250)(13.85) = .000589$$

Asimismo, el límite superior para S^2_{wj} se obtiene por medio de:

$$\text{Pr } \{ S^2_w \mathbf{f} 2n(m-1)s_w^2 / [(2nm-2n-1)]^{1/2} - t_a \}^2 = 1 - a \quad (47)$$

Para el problema que nos ocupa se tiene que:

$$s^2_{w1} = (50)(9)(.0087) / [(500-50-1)^5 - 1.64]^2 = .010244$$

$$s^2_{w2} = (50)(9)(.0432) / [(500-50-1)^5 - 1.64]^2 = .050865$$

Por su parte, los límites inferior y superior de s^2_{uj} se determinan por medio de las siguientes expresiones:

$$[(n-1)s_e^2 / mC^2_1] - [2n(m-1) s^2_w / m[(2nm-2n-1)]^{1/2} - t_{1/2_a}]^2$$

$$[(n-1)s_e^2 / mC^2_0] - [2n(m-1) s^2_w / m[(2nm-2n-1)]^{1/2} + t_{1/2_a}]^2 \quad (48)$$

Siguiendo con nuestro ejemplo, en donde sólo nos interesa conocer el límite superior tenemos que;

$$s^2_{u1} = [(24)(.0492)/(10)(13.85)] - [(50)(9)(.0087)] / \{9[(500-50-1)^5 + 1.96]^2\} = .007714$$

$$s^2_{u2} = [(24)(.0850)/(10)(13.85)] - [(50)(9)(.0432)] / \{9[(500-50-1)^{-5} + 1.96]^2\} = .010699$$

Conforme a los cálculos anteriores, se plantea el siguiente problema de optimización no lineal en la función objetivo y lineal en las restricciones impuestas a los valores de las x_i 's.

$$\text{Max } \mathbf{f} = - [7.8/x_1 + 3/x_2]$$

$$\text{s.a. } .007714 x_1 + .010244 x_2 \leq .000341$$

$$.010699 x_1 + .050865 x_2 \leq .000589 \quad (49)$$

$$0 < x_1 = 1/n \leq 1, \quad 0 < x_2 = 1/nm \leq 1$$

Se observa que de la última relación de las ecuaciones (44) se deduce el número óptimo de selecciones de primera y segunda etapa (n y m), y por lo tanto el total de observaciones (nm) que se deben seleccionar para la realización de la encuesta.

La solución óptima del problema indica que $x_1 = .023404$ y $x_2 = .006657$ por lo que se deben seleccionar un total de $n = 43$ unidades primarias de muestreo (UPM's) y dentro de estas $m = 4$ selecciones de segunda etapa (USM's), lo cual genera un total de $nm = 172$ observaciones elementales (hogares en este caso) a seleccionar para la realización de la encuesta. Si se asume un promedio de 3 personas mayores de 12 años por hogar se tendría un total de 516 personas en edad de trabajar que formarían parte de la investigación.

Otras investigaciones posteriores han estudiado el problema de asignación óptima de la muestra a partir de métodos de programación no lineal a lo cual le han denominado programación convexa por las características de la función objetivo (Kokan y Khan(1967) y Huddleston, Claypool y Hocking(1970). En todas las investigaciones se concluye que no existe solución única y por lo tanto se proponen algoritmos particulares que generan soluciones óptimas bajo ciertas suposiciones que se describen en las propuestas elaboradas.

Los desarrollos analizados muestran que existen alternativas para determinar el número óptimo de selecciones de primera y segunda etapa en las encuestas de propósitos múltiples que deben ser consideradas a fin de mejorar la eficiencia de los diseños de muestra. Asimismo, se demuestra que es posible mejorar los métodos de trabajo actuales aprovechando las capacidades de cómputo y se evidencia la necesidad de generar controles durante la planeación de la encuesta y la ejecución del trabajo de campo que permitan la estimación de los costos que involucra la realización de una encuesta, a fin de poder aplicar algoritmos de cálculo como los que se han analizado.

Además, se hace énfasis en la necesidad de calcular errores de muestreo de los estimadores, así como el efecto de diseño, a fin de disponer de información valiosa para mejorar el cálculo del tamaño de muestra. Esta práctica, sin duda, permitirá mejorar los futuros diseños incrementando la precisión de los estimadores y reduciendo los costos de operación.

IV. CONCLUSIONES

1.- Es claro que una de las preocupaciones principales en la planeación de una encuesta por muestreo probabilístico se orienta a determinar con certeza el número de observaciones necesarias para poder garantizar la precisión y confianza requerida en los estimadores calculados. En este sentido, es importante que las organizaciones ejecutoras estén conscientes que una mala decisión puede limitar la validez de los resultados y en ocasiones llegar a invalidar la investigación completa.

2.- Las encuestas que se realizan en todos los países se pueden considerar de propósitos múltiples. Sin embargo, es práctica común que la determinación del tamaño de muestra se efectúe como si sólo se desearan obtener estimaciones para una sola variable. Posteriormente, se forman dominios de estudios y se analiza la información sin evaluar si las restricciones impuestas por el tamaño de muestra conducen a la pérdida de confiabilidad en las relaciones de causalidad observadas en subpoblaciones de interés y se formulan recomendaciones de política que pueden llegar a carecer de validez estadística.

3.- En muchas ocasiones, se procede como si el único interés de la investigación se centrara en estimar proporciones (p) y posteriormente con los datos generados se forman totales y promedios sin estar conscientes de que el error de muestreo de los estimadores se incrementa por el simple hecho de que para lograr la misma precisión que en el caso de una proporción se requeriría aumentar el tamaño de muestra.

4.- Para el caso de una sola variable la literatura documenta cómo determinar el tamaño de muestra cuando se desea estimar una proporción binomial y no se dispone de un valor que aproxime su varianza. Sin embargo, en la práctica sucede frecuentemente que las variables que se desean estimar tienen una distribución multinomial, por los procedimientos que habitualmente se aplican no resultan adecuado ya que subestima el número de observaciones que se deben seleccionar y no es posible controlar la precisión simultánea de los estimadores.

5.- Para el caso multivariado se han desarrollado procedimientos matemáticos que permiten explorar alternativas y generan soluciones al problema de determinar el tamaño óptimo de muestra y su asignación en los diferentes estratos en que se divide la población.

6.- Cuando se trata de encuestas de propósitos múltiples en donde la muestra se selecciona en diversas fases y se desea obtener el número óptimo de selecciones de primera (n) y segunda etapa (m) es muy difícil obtener una solución única. Sin embargo, existen algunos algoritmos que permiten soluciones bajo ciertas restricciones.

7.- La única manera de saber si se cumplieron las expectativas de confiabilidad de los estimadores definidas en el diseño de la muestra se logra a partir del cálculo de los errores de muestreo y del efecto de diseño, tanto para la variable principal como para todos aquellos indicadores que se analizarán a partir de la información generada.

8.- Para lograr la elaboración de diseños de muestra óptimos es necesario disponer de información sobre la varianza de los estimadores y sobre los costos de operación de las diferentes etapas de una encuesta. En este sentido, los organismos ejecutores debieran aplicar mayores controles en las diversas etapas de planeación y ejecución del trabajo de campo. Esta práctica permitirá mejorar la eficiencia de los diseños, con el consiguiente ahorro de recursos.

REFERENCIAS BIBLIOGRÁFICAS

- Angers, C. (1974). "A Graphical Method to Evaluate Sample Sizes for the Multinomial Distribution". *Technometrics*, vol. 16, No. 3, 469-471.
- Angers, C. (1979). "Sample Size Estimation for Multinomial Populations". *The American Statistician*. Vol. 33, No. 3, 163-164.
- Angers, C. (1989). "Note on Quick Simultaneous Confidence Intervals for Multinomial Proportions". *American Statistical Association*, vol. 43, No.2, 91.
- Bankier, M. D. (1988). "Power Allocation: Determining Sample Size for Subnational Areas". *American Statistician*, vol. 42, No.3, 174-177.
- CEPAL, División de Estadística y Proyecciones Económicas (1988). "Los Errores de Muestreo en las Encuestas Complejas" (documento de trabajo por publicar).
- Cochran, W. G. (1953). "Sampling Technics". John Wiley & Sons, Inc.
- Cowan, C. D. and Malec, D.J. (1988). "Sample Allocation for A Multistage, Multilevel, Multivariate Survey". Bureau of the Census. Fourth Annual Research Conference.
- Chakravarty, I. M. (1955). "On the Problem of Planning a Multistage Survey for Multiple Correlated Characters", *Sankhyā*, vol. 14, 211-216.
- Fitzpatrick, S. and Scott, A. (1987). "Quick Simultaneous Confidence Intervals for Multinomial Proportions". *American Statistical Association*, vol. 82, No. 399, 875-878.
- Ghosh, S. P. (1958). "A note on stratified random sampling with multiple characters", *Calcuta Statisc. Ass. Bull.*, 8, 81-89.
- Goodman, A. Leo. (1965). "On Simultaneous Confidence Intervals for Multinomial Proportions". *Technometrics*, vol. 7, No. 2, 247-254.
- Huddleston, H. F., Claypool, P.L. and Hocking, R.R. (1970). "Optimal Sample Allocation to Strata Using Convex Programming". *Journal of the Royal Statistical Society (Series C)*, vol. 19, 273-278.
- Kish, L. (1965). "Survey Samplig". John Wiley & Sons, Inc.
- Kokan, A. R. (1963). "Optimum Allocation in Multivariate Surveys". *Journal of the Royal Statistical Society (Series A)*, vol. 126, 557-565.

Kokan, A. R. and Khan, S. (1967). "Optimum Allocation in Multivariate Surveys: An Analytical Solution". *Journal of the Royal Statistical Society (Series B)*, vol. 29. 115-125.

Quesenberry, C. P. and Hurst, D. C. (1984). "Large Sample Simultaneous Confidence Intervals for Multinomial Proportions". *Technometrics*, vol. 6, No. 2., 191-195.

Sedransk, J. (1965). "A Double Sampling Scheme for Analytical Surveys" *Journal of the American Statistical Association*. 985-1004.

Sison, C. P. and Glaz, J. (1995). "Simultaneous Confidence Intervals and sample Size determination for Multinomial Proportions". *Journal of the American Statistical Association*, vol. 90, No. 429, 366-369.

Thompson, S. K. (1987). "Sample Size for Estimating Proportions". *The American Statistician*, vol. 41, No.1, 42-46.

Tortora, R. D. (1978). "A Note on Sample Size estimation for Multinomial Populations". *The American Statistician*, vol. 32, No. 3, 100-102.

Waters, J. R. and Chester, A. J. (1987). "Optimal Allocation in Mulivariate, Two Stage Sampling Desings". *The American Statistician*, vol. 41, No. 1, 46-50.

Yates, F. (1953). "Sampling Methods for Census and Surveys". 2nd. Editon, Charles Griffin & Co., London.