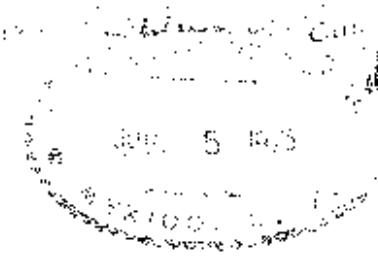
DETERMINACION DE TAMAÑO DE MUESTRA MEDIANTE PROGRAMACION NO-LINEAL

Trabajo de inventigación elaborado por
Angel Rodríguez Vera
Abril de 1975
para optar al grado de Maestro en Estadística de El Colegio de México.



I N D I C E

I N T R O D U C C I O N

En algún momento, durante la planeación de una investigación por - muestreo, es necesario tomar la decisión sobre el número de elementos de- la población que deben ser incluidos en la muestra.

Esta es una decisión importante ya que la calidad de los resulta-- dos de la investigación y el costo de la misma están estrechamente rela-- cionados con el tamaño de muestra que se fije.

Una revisión de la literatura sobre la teoría del muestreo revela- que el tema del tamaño de muestra es tratado de una manera más o menos -- completa únicamente para el muestreo irrestricto aleatorio. Para diseños más complicados el tema se trata más superficialmente y los autores se limitan, la mayoría de las veces, a hacer comentarios o sugerencias sobre - ello, si no es que ignoran el problema por completo. El cálculo del tamaño de muestra plantea una gran variedad de problemas a medida que el diseño se complica. Esto dificulta su tratamiento, ya que, además, a la mayoría de estos problemas no se les ha encontrado una solución teórica ade-- cuada.

Un ejemplo de lo anterior es el cálculo del tamaño de muestra requerido para la estimación de medias o totales de más de una característica o variable, en muestreo estratificado. Es frecuente que la población se estratifique antes de que la muestra sea extraída; de ahí surge la necesidad de fijar el tamaño total de muestra requerido y su distribución en cada uno de los estratos, es decir la afijación de la muestra. La mejor afijación para una característica puede no ser aceptable para otra. Este problema se ha tratado de resolver de diferentes maneras, sin embargo los métodos propuestos hasta ahora no son siempre aplicables y los resultados muchas veces no son satisfactorios.

La finalidad principal de este trabajo es formular la solución del problema anterior por medio de programación no-lineal, y con la aplicación de un algorítmo desarrollado por la Universidad de Wisconsin, presentar algunos resultados que ilustren las ventajas del método.

Con el objeto de darle cierta integridad al contenido de este trabajo, se analiza detenidamente el problema general de la estimación del tamaño de muestra y los factores que pueden influir en él. En los tres primeros Capítulos se presentan los métodos existentes para la estimación del tamaño de muestra en muestreo irrestricto aleatorio y en muestreo estratificado. Este sirve de preámbulo al Capítulo 4 donde se trata el problema multivariado y su solución por métodos de programación no-lineal.

C A P I T U L O 1

ANALISIS DEL PROBLEMA

1.1.- E L P R O B L E M A.

La decisión sobre el tamaño de muestra plantea un conflicto entre los diversos factores que lo afectan; por un lado se deben considerar los recursos económicos de los que se dispone para la investigación y, por -- otro, la calidad de los resultados que se desea obtener. En general, el- aumentar el tamaño de muestra implica un mayor costo, pero una muestra me nor reduce la exactitud de los resultados.

En la mayoría de los casos, la decisión sobre el tamaño de muestra no puede resolverse en forma completamente satisfactoria, debido a que, - por lo regular, no se posee la información suficiente que concilie las -- partes conflictivas, y nos conduzca a lo que podríamos llamar el mejor ta maño de muestra.

Lo anterior debe tenerse presente siempre que se hable del tamaño de muestra, ya que este valor invariablemente va a estar en función de pa rámetros desconocidos de la población, los cuales hay necesidad de esti-- mar. W.G. Cochran [2], quizás para enfatizar este problema, no habla del

cálculo del tamaño de muestra, sino de la estimación del tamaño de muestra.

No debe pensarse, sin embargo, que el problema del tamaño de muestra es irresoluble, o que sólamente se le puede dar una solución burda y poco precisa. En realidad la teoría de muestreo ayuda a tratar inteligentemente el problema y darle una solución adecuada.

1.2.- FACTORES QUE AFECTAN EL TAMAÑO DE MUESTRA .

Se consideran aquí varios factores que, directa o indirectamente, afectan la determinación del tamaño de muestra. Estos factores parecen estar interrelacionados más bien que ser independientes, y su importancia relativa varía enormemente de encuesta a encuesta.

A.- HOMOGENEIDAD O HETEROGENEIDAD DE LA POBLACION. -

Uno de los factores de los que depende el tamaño de muestra requerido es el grado de homogeneidad de la población. Entre más homogénea sea una población se requerirá una muestra más pequeña; por el contrario, entre más heterogénea sea una población el tamaño de muestra deberá ser mayor.

Por homonegeidad de la población se entiende el grado de similitud entre los elementos de la población, con respecto a una característica particular, objeto del estudio, o alguna otra variable correlacionada con dicha característica. Por ejemplo un grupo de personas que tengan las características de ser casados, menores de 40 años y profesar una misma religión, puede ser calificado de homogéneo en un estudio sobre el tamaño de la familia (si el investigador sabe que el número de miembros de una familia está estrechamente relacionado con la edad y la religión). El mismo grupo puede ser calificado de heterogéneo en un estudio sobre ingreso familiar (si el monto del ingreso en el grupo varía grandemente).

El tamaño de muestra requerido para una población heterogénea puede ser abatido mediante una división de la población en estratos, de tal manera que los elementos dentro de cada estrato sean lo más homogéneo posi

ble. Entre más homogéneo sea un estrato requerirá un tamaño de muestra - menor.

B.- ESTIMACIONES PARA SUBDIVISIONES DE LA POBLACION.

Otro elemento importante que se debe tomar en cuenta en la determinación del tamaño de muestra, es el número de categorías o clases en - las que se van agrupar o analizar los datos.

Investigadores inexpertos frecuentemente se sorprenden de la rapidez con la que el tamaño de muestra requerido aumenta cuando se forman - subgrupos o dominios de la población. Aún cuando una muestra sea absolutamente adecuada para la tabulación principal, el tamaño de muestra se hace rápidamente inadecuado cuando se preparan tabulaciones más desglozadas. Por ejemplo, una muestra de 1000 familias puede resultar adecuada para -- investigar el porcentaje de familias que poseen casa propia. Si el 30 -- por ciento de las familias de la población tienen casa propia, se obtendrá, en promedio, ese mismo porcentaje en la muestra es decir 300. Pero- el investigador está además interesado en conocer otras características - de las personas que tienen casa propia, por ejemplo si son residentes rurales o urbanos, qué grupos de ingresos representan y cuántos de ellos -- son campesinos. Si en la muestra la clasificación rural-urbana produjera 200 residentes urbanos y 100 rurales, y si el grupo rural fuese clasificado en 6 grupos de ingreso, el grupo más alto de ingreso podría contener - únicamente 5 casos. Para determinar si en el grupo más alto de ingreso - los residentes rurales son campesinos o no, se tendría que hacer la estimación con únicamente 5 elementos.

Si la frecuencia en cada subclase es tratada como una muestra de- una subpoblación o dominio (como en el anterior ejemplo, donde el investigador desea describir las características de ocupación de los residentes- rurales, con los ingresos más altos y que tienen casa propia), los 5 ca-- sos en la subclase mencionada constituyen la muestra de este dominio, y -

la bondad de las estimaciones que se hagan de esta subpoblación están en - función de este número, i.e. 5. Podemos considerar esta subclase independiente, ya que la muestra tomada ahí representa exclusivamente a la subpoblación definida por ella y no a la población total, representada por la - muestra de 1000 familias.

Ahora bien, las subclases que representan meramente una porción de un grupo más grande, y no pueden considerarse como una subpoblación independiente para los propósitos de análisis, no deben ser tratados como independientes. En el ejemplo anterior el tamaño total de la muestra es 1000; si los 5 casos son expresados como un porcentaje (o alguna otra medida) -- con base en 1000, la muestra total será 1000 y no 5.

De lo anterior se puede concluir que el tamaño de muestra total escogida, debe ser lo suficientemente grande para asegurar tamaños de muestra que representen bien a las subclases independientes más pequeñas.

C.- PRECISION Y CONFIANZA DE LA ESTIMACION.

¿Qué tanto puede variar una estimación del verdadero valor de la - población (desconocido) y todavía ser aceptable? supongamos que se quiere saber qué porcentaje de una población de mujeres trabajan. ¿Será suficiente para el investigador demostrar que dicho porcentaje es menor del 20 por ciento? o ¿tendrá que probar que es menor del 20 por ciento pero mayor del 18 por ciento?. La respuesta a este tipo de preguntas depende del objetivo particular de la investigación. El investigador puede permitir un 20 ó un 2 por ciento de error, dependiendo de la utilización que vaya a dar a- los resultados.

En ocasiones, el muestrista está en condiciones de hacer sugerencias con respecto a la exactitud de las estimaciones, debido a sus conocimientos de los requerimientos técnicos de la investigación. Aún más, es- frecuente que la precisión requerida de las estimaciones dependa más de - dichos requerimientos, que de la opinión de la persona que patrocina el -

estudio. Supongamos que una persona desea predecir cuál de dos candidatos va a ganar unas elecciones y no especifica la precisión que desea del porcentaje que se va a estimar. Es entonces el muestrista quien debe tomar la decisión. Si se espera que la votación sea muy cerrada, el muestrista no puede tolerar un error muy grande, si quiere estar relativamente seguro de quién ganará las elecciones. Por ejemplo, si se estima que el 55 por ciento de los votos favorecerá a un candidato, el error permisible debe ser menor del 5 por ciento, de otra manera el resultado de las elecciones no se puede anticipar. Si en vez de 55 por ciento obtiene un 60 por ciento de los votos para un candidato, un error menor o igual a 9 por ciento puede ser tolerado. Diremos entonces, que el error permisible (precisión) de una estimación, es la máxima desviación del verdadero valor, que el investigador está dispuesto a aceptar. Este valor permitirá al muestrista resolver satisfactoriamente cuestiones inherentes a aspectos técnicos de la encuesta.

Ahora bien, aparte de un censo completo, no hay nada que asegure completamente que la estimación no excede del error deseado. Se deberá, entonces, decidir el grado de seguridad o probabilidad de tener un error menor o igual al permisible, de acuerdo con la situación particular. Si el proyecto es importante y las facilidades son apropiadas y suficientes, no hay razón para abstenerse de fijar una seguridad o confianza grande. Si por el contrario, el tiempo y los fondos son limitados, se deberá pensar en conformarse con una menor confianza, dentro de ciertos límites adecuados, para las estimaciones.

D.- RECURSOS DISPONIBLES.

El tamaño de muestra debe ser necesariamente consistente con los recursos disponibles para realizar la investigación. Para ello se requiere una estimación del costo, tiempo y materiales que serán necesarios para un determinado tamaño de muestra. Puede ser que un tamaño de muestra que-

se ha fijado y que satisface todas las demás factores, no pueda ser mantenido por falta de recursos; esto conducirá a tomar la decisión de disminuir el tamaño de muestra (y reducir la precisión) o posponer la investigación hasta que se disponga de los recursos necesarios.

E.- NUMERO DE CARACTERISTICAS QUE SE INVESTIGAN.

Es común que en una encuesta no se investigue únicamente una característica o variable, sino varias de ellas. Cuando el número de variables es grande y se especifica un grado de precisión para cada una, los cálculos conducen a una serie de valores diferentes y por lo tanto conflictivos, de tamaños de muestra. Es entonces cuando se debe contar con algún método para conciliar estos valores.

C A P I T U L O II

TAMAÑO DE MUESTRA EN EL MUESTREO IRRESTRICTO ALEATORIO

2.1.- MUESTREO IRRESTRICTO ALEATORIO .

Con la expresión 'muestreo irrestricto aleatorio' se denomina al diseño que utiliza el procedimiento más elemental que existe para seleccionar una muestra. Todos los demás procedimientos de selección pueden considerarse como una variación de él, para proporcionar más eficiencia, economía o comodidad en el diseño de la muestra.

Así mismo, la estimación del tamaño de muestra en el muestreo irrestricto aleatorio, proporciona un punto de partida para tratar el problema en situaciones y diseños más complejos.

2.2.- TAMAÑO DE MUESTRA PARA PROPORCIONES .

Se tiene una población dicotómica, es decir los elementos pertenecientes a esta población pueden clasificarse en dos clases mutuamente exclusivas y exhaustivas, digamos A y A'. Si se especifica un margen de error d en la estimación de la proporción P, del número de elementos en la clase A, y un riesgo α que representa la probabilidad de que el error real sea más grande que d, podemos formular lo siguiente:

$$P_r \ (\ | \ p - P \ | \ \geq \ d \ ) \ = \ \alpha \qquad (2.1)$$

donde p es un estimador de P.

Si el muestreo es irrestricto aleatorio y el tamaño de muestra es suficientemente grande, p se puede suponer distribuida normalmente y entonces (W.G. Cochran [2] pág. 50),

$$\sigma_p^2 \ = \ \left(\frac{N - n}{N - 1} \right) \ \left(\frac{P \ (1 - P)}{n} \right) \qquad (2.2)$$

donde N es el tamaño total de la población y n el tamaño de la muestra que se selecciona de ella.

(2.1) puede expresarse también como:

$$P_r \ (P - d \ < \ p \ < \ P + d) \ = \ 1 - \alpha \qquad (2.3)$$

donde $p \sim N (P, \sigma_p^2)$

y por lo tanto $\dfrac{p - P}{\sigma_p} \sim N (0,1)$

la expresión (2.3) puede entonces reescribirse como:

$$P_r \left(- \frac{d}{\sigma_p} < \frac{p - P}{\sigma_p} < \frac{d}{\sigma_p} \right) = 1 - \alpha$$

Si $t = d/\sigma_p$ se obtiene la fórmula que relaciona a n con un grado deseable de precisión, esto es,

$$d \ = \ t \ \sigma_p = t \sqrt{\left(\frac{N - n}{N - 1} \right) \ \left(\frac{P \ (1 - P)}{n} \right)}$$

donde t es la abscisa a la curva normal con media 0 y varianza 1, --

que corresponde a una área $\alpha/2$ en cada cola.

Al despejar n se tiene:

$$n = \frac{\dfrac{t^2 P (1 - P)}{d^2}}{1 + \dfrac{1}{N}\left(\dfrac{t^2 P (1 - P)}{d^2} - 1\right)} \qquad (2.4)$$

Para efectos prácticos, una primera aproximación para el tamaño de muestra sería:

$$n' = \frac{t^2 P (1 - P)}{d^2} \qquad (2.5)$$

Si n'/N es despreciable, n' es una aproximación satisfactoria de n. Si no es así, el valor exacto de n será:

$$n = \frac{n'}{1 + (n' - 1)/N} = \frac{n'}{1 + (n'/N)}$$

Algunos autores llaman a esta expresión 'corrección para poblaciones finitas'.

A continuación se dan algunos ejemplos para ilustrar el uso de estas fórmulas.

Volvamos a la encuesta para predecir cuál de dos candidatos va a ganar unas elecciones, por lo que se desea estimar el porcentaje P de votos a favor de uno de los candidatos. Los resultados anteriores nos permiten determinar el tamaño de muestra necesario, si utilizamos un muestreo irrestricto aleatorio.

La fórmula (2.5), que nos da el tamaño de muestra adecuado, está-

en función precisamente del parámetro P que se desea estimar.

Si el investigador no tiene ninguna evidencia de cuál candidato va a ganar y por qué margen, lo mejor sería situarse conservadoramente y tomar el valor máximo de n que la fórmula (2.5) puede proporcionar; esto sucede cuando P = 0.5. El tamaño de n que se obtenga con ese valor de P garantizará, cuando menos, la precisión que se haya fijado.

Por ejemplo, si d = 0.05 y α = 0.05, entonces t = 1.96. Al sustituir estos valores en (2.5) con P = 0.5, y suponiendo que N es grande, por lo que no hay necesidad de aplicar la corrección para poblaciones finitas, se obtiene:

$$n = \frac{(1.96)^2 (0.5) (0.5)}{(0.05)^2} = 384$$

Este valor de n asegura que el estimador p tomará un valor dentro del intervalo (P ± .05) con una probabilidad de cuando menos .95.

Son muchas las ocasiones en las cuales se puede tener evidencia a-priori del valor de P. Esta evidencia puede tener diversos orígenes, puede provenir de un estudio anterior semejante, de una encuesta piloto o bien deducirse de la naturaleza misma de la encuesta que se va a realizar.

Por ejemplo, supongamos que se quiere seleccionar una muestra para *estimar el porcentaje de personas que son* profesionistas en una cierta población. El investigador sabe de antemano que la proporción de profesionistas en esa población es pequeña y nos dice que no puede ser mayor de -- .20 Esta información es valiosa para el cálculo del tamaño de muestra, ya que con este dato adicional se puede obtener un tamaño de muestra más adecuado para este problema en particular.

Si despreciamos la corrección para poblaciones finitas y fijamos-- P = 0.20, d = 0.05, α = 0.05 y t = 1.96, el tamaño de muestra

requerido es:

$$n = \frac{(1.96)^2 \ (.2)\ (.8)}{(.05)^2} = 246$$

Al comparar este último valor con el obtenido en el ejemplo anterior, es decir 384, resulta evidente la ganancia obtenida en el tamaño de muestra debido a un conocimiento a priori, aunque burdo, del parámetro P.

La precisión de la estimación del parámetro P puede plantearse de otra manera. Es evidente que el tamaño de muestra está en función de la varianza del estimador, por lo tanto se puede fijar un valor V tal que:

$$\sigma_p^2 = V \qquad (2.6)$$

O bien.

$$\left(\frac{N - n}{n - 1}\right) \left(\frac{P\,(1 - P)}{n}\right) = V$$

de donde se obtiene que el tamaño de muestra requerido para satisfacer (2.6) es:

$$n' = \frac{P\,(1 - P)}{V}$$

O bien si N no es muy grande

$$n = \frac{n'}{1 + (n'/N)}$$

Cuando la población puede clasificarse en más de dos clases, digamos $A_1, A_2, \ldots, A_K$, y desean estimarse las proporciones correspondientes $P_1, P_2, \ldots, P_K$, la estimación del tamaño de muestra no representa un problema diferente al anterior.

Si se desea calcular el tamaño de muestra requerido para estimar - la proporción P_i, ese tamaño deberá calcularse según las fórmulas deducidas anteriormente, ya que en realidad la clasificación i divide a la población en dos partes, que son A_i y todo lo que no es A_i, digamos A'_i.- De esta manera se obtiene un tamaño de muestra necesario para estimar cada P_i, según la precisión y confianza fijados para cada una de ellas. El tamaño de muestra que satisface todas las clasificaciones es:

$$\text{Máx} \quad (n_1, n_2, \cdots, n_k)$$

donde:
$$n_i = \frac{n'_i}{1 + (n'_i/N)}$$

y
$$n'_i = \frac{t_i^2 P_i (1 - P_i)}{d_i^2}$$

2.3.- TAMAÑO DE MUESTRA PARA MEDIAS Y TOTALES.

Otro caso importante en el cálculo del tamaño de muestra, es el de la estimación de medias de la población, el cual es análogo a la estimación de proporciones descrito en la sección anterior.

El problema es ahora calcular el tamaño de muestra necesario para estimar la media de la población, dentro de ciertos límites de error y con una probabilidad dada. Podemos expresar lo anterior de la siguiente manera:

$$P_r (| \bar{y} - \bar{Y} | \geq d) = \alpha$$

Donde d es la precisión, representada por un margen de error que se ha fijado, y α es la probabilidad de que la estimación se exceda de ese margen de error. Si suponemos que $\bar{y}$ se distribuye normalmente su varianza es (G. W. Cochran [2] pág. 23)

$$\sigma_{\bar{y}}^2 = \left(\frac{N - n}{N}\right) \left(\frac{S^2}{n}\right)$$

donde S^2 es la varianza por elemento.

Siguiendo el mismo procedimiento que se utilizó para obtener la -- fórmula para proporciones, encontramos que:

$$n' = \left(\frac{t\,S}{d}\right)^2$$

y

$$n = \frac{n'}{1 + (n'/N)}$$

Si se requiere la fórmula de n en función de la varianza del estimador, V, entonces.

$$n' = \frac{S^2}{V}$$

En algunas ocasiones es útil considerar una medida relativa de la- variación de una variable, en lugar de una medida absoluta. Las medidas - absolutas, como la desviación estándar, se expresan en las mismas unidades que la variable, y esto causa ciertas dificultades, sobre todo al hacer -- comparaciones. Una medida relativa de la variación de una variable es el- 'coeficiente de variación'. En él, la unidad de medida se cancela al dividir entre la media:

$$C = \frac{S}{\bar{Y}}$$, que se estima mediante $$c = \frac{s}{\bar{y}}$$

Similarmente, el coeficiente de variación de la media ($\bar{y}$) es:

$$CV(\bar{y}) = \frac{S_{\bar{y}}}{\bar{Y}}$$ estimado por $$cv(\bar{y}) = \frac{s_{\bar{y}}}{\bar{y}}$$

Muchas veces resulta útil obtener la fórmula de n en términos de varianzas relativas (L. Kish [8] pág. 50). Sabemos que:

$$CV^2(\bar{y}) = \left(1 - \frac{n}{N}\right)\left(\frac{C^2}{n}\right)$$

De donde:

$$n' = \frac{C^2}{CV^2(\bar{y})} \quad y \quad n = \frac{n'}{1 + (n'/N)}$$

Estas fórmulas se aplican directamente a totales del tipo $N\bar{y}$, ya que las varianzas relativas son las mismas para el total $N\bar{y}$ que para la media $\bar{y}$.

Como se menciona en el primer Capítulo, siempre nos encontraremos que n (el tamaño de la muestra) está en función de un parámetro de la población que desconocemos. En las fórmulas anteriores ese parámetro es S^2.

L. Kish [6] explica ampliamente este problema en los comentarios siguientes:

En la práctica se desconoce S^2, y su valor se tiene que estimar o conjeturar. ¿Cuáles son las fuentes de esas conjeturas?.

1.- Debemos buscar datos de encuestas anteriores de variables semejantes, o pedir consejo a un estadístico experto en encuestas, que tenga conocimiento de encuestas anteriores y habilidad para descubrir sus aspectos sobresalientes. Para obtener los datos relevantes, el estadístico podrá preguntar a los especialistas en la materia que corresponda al estudio. Al apoyarnos en ese conocimiento, podemos construir un modelo de la distribución de la población, determinar su forma y sus límites probables y deducir el valor de S^2.

2.- Si conocemos la varianza Var $(\bar{y})$ de un muestreo irrestricto anterior de tamaño n^*, entonces podemos utilizar $S^2 = (\text{Var}(\bar{y}))(n^*/(1 - n^*/N))$. Si la muestra no es irrestricta aleatoria, puede utilizarse el 'efecto de diseño' para ajustar la varianza.

3.- A menudo, en lugar de S es más fácil suponer el valor de $C = S/\bar{Y}$, porque C es menos variable que S; por tanto, es más fácil tomar datos de los resultados de variables semejantes. Con una estimación de C y también de $\bar{Y}$ podemos estimar $S = C\bar{Y}$.

4.- Para diseñar eficientemente una muestra grande en un campo desconocido, se puede realizar un estudio piloto anterior a la encuesta, con objeto de obtener información para diseñar la encuesta. Pero la mayoría de los estudios o bien son demasiado pequeños, o bien se tienen que hacer demasiado rápidamente para resistir un estudio piloto suficientemente grande que produzca estimaciones útiles de S^2. Si el estudio piloto es muy pequeño, sus resultados son poco útiles, puesto que son menos confiables que las conjeturas de expertos, las cuales podemos obtener sin él.

A veces, el tamaño de la muestra puede ajustarse más estrechamente a las demandas de la encuesta. Este ajuste será mejor si se tiene flexibilidad en el diseño. Primero, se recolecta una muestra básica de un tamaño razonablemente mínimo para satisfacer las demandas. A continuación, se calculan los resultados y, si las demandas no se satisfacen, se vuelve a tomar una muestra suplementaria del tamaño requerido. Este procedimiento en dos pasos puede utilizarse para obtener o bien una varianza que se desea, o bien un tamaño de muestra. Sin embargo, no se puede utilizar en encuestas con calendarios rígidos de tiempo que impidan aplicar este procedimiento de dos pasos.

Naturalmente los valores supuestos con respecto a S^2 están sujetos a error, y la varianza definitiva de la media de la muestra puede ser menor o mayor de lo que se haya planeado. Pero estos errores no afectan -

la validez de las varianzas que se calculan a partir de los valores reales de la muestra, que no son influidos en absoluto por los valores que se imputan al principio a S^2.

Además, los errores en las estimaciones de S^2 suelen ser excedidos por el margen de ignorancia con respecto a varios temas relacionados. Primero: a menudo sabemos aún menos acerca de los factores de costo que acerca de S^2. Segundo: el enunciado que se hace de una varianza como adecuada o deseada generalmente está sujeto a mayores vaguedades de lo que está S^2. Esto es cierto sobre todo cuando en la encuesta hay varios objetivos, con demandas que están en conflicto con el tamaño de muestra deseable. Tercero: nuestro conocimiento acerca de los efectos de los errores no de muestreo no es, en general, tan adecuado como el que tenemos acerca de S^2.

2.4.- TAMAÑO DE MUESTRA PARA ESTIMACIONES DE SUBDIVISIONES DE LA POBLACION.

Como se comenta en el primer capítulo, es muy frecuente que se requieran estimaciones no solamente para la población como un todo, sino también para ciertas subdivisiones de ella.

Si las subdivisiones pueden ser identificadas a priori es decir que se sabe a que subdivisión pertenece cada elemento de la población antes de tomar la muestra, entonces se puede hacer el cálculo de n para cada subdivisión. Por ejemplo, supongamos que se quiere estimar el rendimiento promedio de maíz por hectárea de todo un país, pero además se desea conocer el rendimiento promedio en cada estado. Si la media de cada estado o subdivisión se quiere estimar con una varianza específica V, entonces para la i-ésima subdivisión se tiene que $n_i = S_i^2/V$ (sin tomar en cuenta la corrección para poblaciones finitas), y el tamaño de muestra total requerido es:

$$n = \sum_i S_i^2/V$$

Ahora bien, si las subdivisiones de una población de personas representan clasificaciones por variables tales como edad, sexo, ingreso o años de escolaridad, la subdivisión a la cual una persona pertenece no es conocida hasta que la muestra ha sido tomada. En estas condiciones el problema se complica, ya que no se puede garantizar un determinado tamaño de muestra para cada subdivisión.

Si se conocen las proporciones π_i del número de elementos de cada subdivisión con respecto al total, entonces es todavía posible estimar el tamaño del total de la muestra. Si se selecciona una muestra irrestricta aleatoria de tamaño n, el tamaño de muestra esperado en la subdivisión i es $n\pi_i$. La varianza promedio de la media de esta subdivisión es:

$$V(\bar{y}_i) = E\left(\frac{S_i^2}{n_i}\right) \doteq \frac{S_i^2}{n\pi_i}$$

Si $n\pi_i$ es grande. Por lo tanto se requerirá $n \doteq S_i^2 / \pi_i V$ para hacer $V(\bar{y}_i) = V$. Si esto vale para cada subdivisión, entonces

$$n \doteq \text{máx}\left(\frac{S_i^2}{\pi_i V}\right)$$

T. Garza y J. A. Coronel [11] tratan este problema de una manera interesante, la cual se expone a continuación.

Limitan el problema a poblaciones finitas y establecen un conjunto de condiciones que debe cumplir cada uno de los números n_i que resultan en cada subdivisión al extraer la muestra. Estas condiciones las fija el investigador en términos de las necesidades de información que requiera en cada subdivisión, y generalmente pueden establecerse mediante la colección de desigualdades

$$n_1 \geq d_1,\ n_2 \geq d_2,\ \ldots,\ n_k \geq d_k$$

donde las d_i son constantes, no negativas y arbitrarias.

La naturaleza aleatoria del experimento no permite asegurar que se cumplirán las anteriores desigualdades, pero se puede, en cambio, prescribir una probabilidad mínima de que suceda, y la magnitud de dicha probabilidad, que en general dependerá del tamaño de la muestra (y, por supuesto, de las d_i), será una medida de la bondad esperada de la muestra para los fines del investigador.

En las condiciones anteriores, el problema puede plantearse de la siguiente manera: determinar un número n tal que satisfaga la desigualdad.

$$P_r\,(n_1 \geq d_1,\ \ n_2 \geq d_2,\ \ldots,\ n_k \geq d_k) \ \geq\ \beta \qquad (2.7)$$

Donde β es un número positivo y menor que 1, que se fija arbitrariamente en atención a las necesidades de la investigación: un valor pequeño de β dará lugar a un tamaño de muestra relativamente pequeño, pero quequizá no garantizará debidamente que las subdivisiones queden representadas en la muestra, en tanto que una β demasiado grande garantizará representatividad, pero llevará con seguridad a un tamaño de muestra excesivamente alto.

La solución que se encuentra a este problema así planteado, es la siguiente:

$$1 - \sum_{i=1}^{k} P_r\,(n_i < d_i) \leq P_r\,(n_1 \geq d_1\ ,\ n_2 \geq d_2,\ldots,n_k \geq d_k) \leq \prod_{i=1}^{k} P_r\,(n_i \geq d_i)$$

Es decir se establece una cota inferior y otra superior para la probabilidad (2.7). Ambas cotas aparecen sólo en términos de las probabilidades marginales correspondientes $P_r\,(n_i \geq d_i)$ y $P_r\,(n_i < d_i)$, que son complemento uno de otra y fáciles de calcular. En efecto, se tiene--

que:

$$P_r\,(n_i \geq d_i) = \sum_{j=d_i}^{n} \frac{\binom{N_i}{j}\binom{N - N_i}{n - j}}{\binom{N}{n}} \qquad (2.8)$$

donde N_i es el número total de elementos en la subdivisión i y $N = \sum_i N_i$.

Cuando las A_i son grandes, la distribución hipergeométrica tiende a una distribución binomial con parámetros $(n, N_i/N)$; puede entonces usarse esta última o bien la distribución normal respectiva para evaluar la suma (2.8) sin pérdida de precisión en los cálculos.

Por ejemplo, considérese el problema de extraer una muestra aleatoria de la población del Distrito Federal, de manera que se tengan representadas las siguientes características simultáneamente: sexo, edad, sector de ocupación, condición migratoria y tamaño de la localidad de residencia. Si se toman dos categorías de sexo, 9 grupos de edad, 5 sectores de ocupación, 2 condiciones migratorias migrantes y no migrantes) y 3 tamaños de localidad de residencia (1 - 2 500, 2501 - 20 000 y 20 001 en adelante), tenemos un total de 540 subdivisiones, y se trata, entonces, de determinar un tamaño de muestra que garantice una representatividad adecuada de las mismas dentro de la muestra.

Puede observarse, *a priori*, que un problema de esta índole va a requerir de una muestra muy grande, en virtud del alto nivel de detalle que se estipula.

Este ejemplo se presentó en la práctica y el marco muestral utilizado fué la información del Censo de Población de 1960, de ahí se obtuvieron estimaciones para los valores de N_i correspondientes a cada celda, y con base en ellas se efectuaron los cálculos para determinar el tamaño de muestra que sería necesario extraer del Censo de 1970 a fin de lograr la

representación deseada en cada una de las 540 celdas definidas anteriormente.

Las sumas (2.8) se calcularon utilizando la aproximación normal a la distribución hipergeométrica:

$$P_r\ (n_i \geq d_i) \doteq \frac{1}{\sqrt{2\pi}} \int_{a_i}^{\infty} e^{-y^2/2}\, dy$$

Donde $a_i = (d_i - n N_i/N - 0.5/ \sqrt{(nN_i/N)\ (1 - n N_i/N)}$.

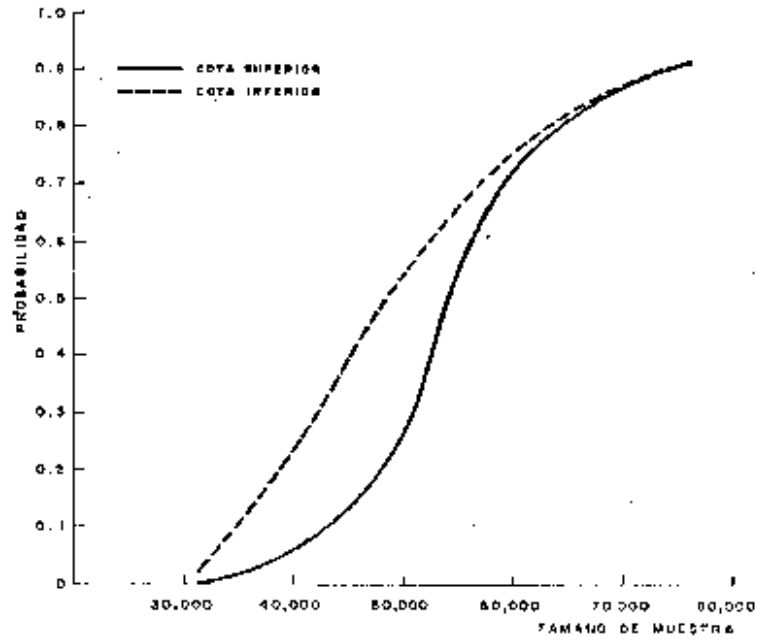
Se consideraron las siguientes dos posibilidades para las d_i :

a) Se requiere que la muestra contenga, en la subdivisión i-ésima, el mínimo de las cantidades (50, 0.02 x N_i), en el caso en que dicha subdivisión contenga al menos 300 individuos en el Censo de 1960. Si la subdivisión tenía menos de 300, no se le impone condición alguna.

b) Se requiere lo mismo que en a) para las subdivisiones que en 1960 tenían al menos 300, y además el 8% del total si éste era inferior a 300 en 1960.

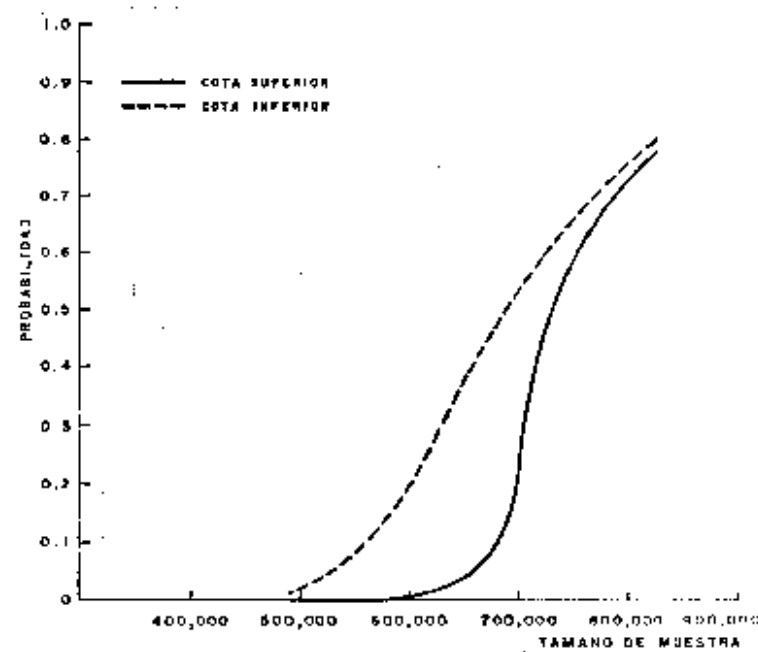
Nótese que la condición b) es mucho más restrictiva que la a), y esto se refleja, como era de esperar, en el tamaño de la muestra.

Los cálculos necesarios se efectuaron haciendo uso de un computador electrónico, los resultados se presentan en las gráficas 1 y 2, donde aparecen los tamaños de muestra correspondientes a diferentes probabilidades de garantizar las condiciones a) y b), respectivamente. Es interesante observar la eficiencia de las cotas propuestas en este trabajo, pues a partir de aproximadamente 0.8 se ve que de hecho coinciden los dos casos.



Gráfica 1

COTAS SUPERIOR E INFERIOR A LA PROBABILIDAD DE SATISFACER LOS REQUISITOS *a)*



Gráfica 2

COTAS SUPERIOR E INFERIOR A LA PROBABILIDAD DE SATISFACER LOS REQUISITOS *b)*

Fuente: T. Garza H. y J.A. Cornel. Demografía y Economía 10.

C A P I T U L O III

TAMAÑO DE MUESTRA EN MUESTREO ESTRATIFICADO

3.1.- MUESTREO ESTRATIFICADO.

El muestreo estratificado se usa muy frecuentemente, debido a que tiene varias ventajas importantes que conducen a diseños más eficientes.- En términos generales, la estratificación consiste en clasificar a una población en dos o más grupos o clases llamados estratos. De cada estrato se extrae entonces una muestra independiente. El número de elementos seleccionados dentro de cada estrato puede ser proporcional o desproporcional al tamaño del estrato. Si cada estrato no se representa en la muestra con la misma proporción que tiene en la población, el estimador por estrato deberá ponderarse para obtener el estimador de toda la población.

3.2.- ESTIMACION DE n PARA MEDIAS Y TOTALES

El estimador usado en muestreo estratificado para estimar una media $\bar{Y}$, es:

$$\bar{y} = \sum_{h=1}^{L} N_h \bar{y}_h / N$$

Donde: $\bar{y}_h$ es un estimador de la media del estrato h, es decir $\bar{Y}_h$.

N_h es el número de elementos en el estrato h, y $N = \Sigma N_h$.

La varianza del estimador $\bar{y}$, si $\bar{y}_h$ es un estimador insesgado de $\bar{Y}_h$, es (W.G. Cochran [2] pág. 91).

$$V(\bar{y}) = \frac{1}{n} \Sigma \frac{W_h^2 S_h^2}{w_h} - \frac{1}{N} \Sigma W_h S_h^2 \qquad (3.1)$$

donde S_h^2 es la varianza por elemento del estrato h, $W_h = N_h/N$ y $w_h = n_h/n$.

Si se desea que n sea tal que.

$$V(\bar{y}) = V$$

Se obtiene, como fórmula general para n,

$$n = \frac{\Sigma \frac{W_h^2 S_h^2}{w_h}}{V + \frac{1}{N} \Sigma W_h S_h^2} \qquad (3.2)$$

Si se ignora la corrección para poblaciones finitas se tiene, como una primera aproximación,

$$n' = \frac{1}{V} \Sigma \frac{W_h^2 S_h^2}{w_h} \qquad (3.3)$$

Si n'/N no es despreciable, se puede calcular n mediante

$$n = \frac{n'}{1 + \frac{1}{NV} \sum W_h S_h^2} \qquad (3.4)$$

3.3.- AFIJACION OPTIMA.

En las fórmulas presentadas anteriormente no se ha considerado el costo como un factor que afecta al tamaño de la muestra. El costo juega un papel importante en el muestreo estratificado, ya que plantea un conflicto en la manera como se reparta o afije la muestra en cada uno de los estratos. Si el costo por elemento muestrado es muy alto en un determinado estrato, lo conveniente es extraer de él la menor muestra posible; inversamente, si es barato, conviene que la muestra en ese estrato sea grande.

La fórmula (3.2) proporciona el tamaño total n de la muestra que se debe seleccionar en un diseño estratificado, para satisfacer una varianza determinada del estimador. Este valor de n está en función de $w_h = n_h/n$, es decir de manera cómo la muestra n se reparte en cada uno de los estratos. Por lo tanto, los valores de n_h se pueden seleccionar de tal manera que satisfagan ciertas condiciones, por ejemplo que la varianza $V(\bar{y})$ sea mínima para un costo especificado, o bien que se minimice el costo para un valor determinado de $V(\bar{y})$.

La función de costo más simple es de la forma:

$$\text{costo} = C = C_0 + \sum c_h n_h \qquad (3.5)$$

Es decir, dentro de cada estrato el costo es proporcional al tamaño de la muestra, pero el costo por unidad muestreada puede variar de estrato a estrato. El término c_0 representa un costo fijo independiente de los estratos y del tamaño de la muestra.

Ciertos tipos de muestreo pueden tener costos que no están linealmente relacionados al número de unidades que tiene cada estrato en la muestra. Una función de costo más general sería,

$$C = c_o + \sum_h c_h n_h^q \qquad (3.6)$$

Por ejemplo, cuando el costo principal de muestreo es el de viajar entre unidades, la relación (3.6) con $q < 1$ (por ejemplo, $q = .5$) podría ser más realista, y c_h sería el costo de viajar entre cada unidad. En este trabajo únicamente se considerará la función lineal de costo (3.5).

Considérese el problema de minimizar

$$V(\bar{y}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{N_h}$$

Sujeta a la restricción

$$c_1 n_1 + c_2 n_2 + \cdots + c_L n_L = C - c_o$$

Con el uso del método de multiplicadores de Lagrange, se deben encontrar las n_h y el multiplicador λ que minimicen

$$V(\bar{y}) + \lambda \left(\Sigma c_h n_h - C + c_o \right)$$

$$= \Sigma \frac{W_h^2 S_h^2}{n_h} - \Sigma \frac{W_h^2 S_h^2}{N_h} + \lambda \left(c_1 n_1 + c_2 n_2 + \cdots + c_L n_L - C + c_o\right)$$

Al diferenciar con respecto a n_h se obtiene el conjunto de ecuaciones:

$$- \frac{W_h^2 S_h^2}{n_h^2} + \lambda c_h = 0 \quad (h = 1, 2, \cdots L)$$

O sea
$$n_h \sqrt{\lambda} = \frac{W_h S_h}{\sqrt{c_h}} \qquad (3.7)$$

Si se suma sobre todos los estratos, se obtiene

$$n \sqrt{\lambda} = \Sigma \frac{W_h S_h}{\sqrt{c_h}} \qquad (3.8)$$

Finalmente, la razón de (3.7) a (3.8) da:

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\Sigma (W_h S_h / \sqrt{c_h})} = \frac{N_h S_h / \sqrt{c_h}}{\Sigma (N_h S_h / \sqrt{c_h})} \qquad (3.9)$$

Este resultado nos lleva a establecer las siguientes reglas de conducta:

En un estrato determinado extraer una muestra más grande si.

1. El estrato es más grande
2. El estrato es más variable internamente
3. El costo por unidad muestral es más barato.

La ecuación (3.9) da el valor de n_h en términos de n, pero no se conoce todavía el valor de n. La solución depende de si el tamaño de la muestra se obtiene fijando un costo total C, o bien se fija una varianza V para $\bar{y}$. Si el costo es fijo, se sustituyen los valores óptimos de n_h en la función de costo (3.5), y se resuelve para n. Esto da:

$$n = \frac{(C - c_0) \Sigma (N_h S_h / \sqrt{c_h})}{\Sigma (N_h S_h \sqrt{c_h})} \qquad (3.10)$$

Si V es fija, se sustituyen las n_h óptimas en la fórmula para $V(\bar{y})$. De esta manera se encuentra que:

$$n = \frac{\left(\sum W_h S_h \sqrt{c_h} \right) \sum W_h S_h / \sqrt{c_h}}{V + (1/N) \sum W_h S_h^2} \qquad (3.11)$$

Un caso especial se presenta cuando $c_h = c'$, es decir el costo por unidad es el mismo en todos los estratos. En este caso la función de costo es $C = c_o + c'n$. Si se sustituye c_h por c' en (3.9), se obtiene:

$$\frac{n_h}{n} = \frac{W_h S_h}{\sum W_h S_h} = \frac{N_h S_h}{\sum N_h S_h} \qquad (3.12)$$

Esta manera de repartir o afijar la muestra se denomina afijación óptima de Neyman, por ser J. Neyman quien dió este resultado.

La fórmula para n cuando se usa la afijación óptima de Neyman, y además se desea estimar a $\bar{Y}$ con una varianza V y con los mismos costos entre estratos, se obtiene sustituyendo c' por c_h en (3.11), es decir:

$$n = \frac{\left(\sum W_h S_h \right)^2}{V + \frac{1}{N} \sum W_h S_h^2} \qquad (3.13)$$

Las fórmulas para n cuando se quiere estimar el total Y, se obtienen inmediatamente a partir de los anteriores resultados. Por ejemplo si V' es un valor fijo deseable de $V(Y)$, el tamaño de muestra requerido, si se usa afijación de Neyman y los costos son los mismos para cada estrato, es:

$$n = \frac{\left(\sum N_h S_h \right)^2}{V' + \sum N_h S_h^2}$$

3.4.- ESTIMACION DE n PARA PROPORCIONES

El estimador p de la proporción P puede ser una media, por ejemplo si se define a la variable y como

$$y = \begin{cases} 1 & \text{si el elemento tiene cierta característica} \\ 0 & \text{si no la tiene.} \end{cases}$$

Un estimador de la proporción P del número de elementos de la población que poseen cierta característica, construído con base en una muestra de tamaño n, es:

$$p = \frac{\sum_{i=1}^{n} y_i}{n}$$

Por lo tanto podemos utilizar los resultados obtenidos en las secciones anteriores, para encontrar las fórmulas para n correspondientes a proporciones. Para esto bastará sustituir S_h^2 por la forma particular que tiene para proporciones, es decir (W.G. Cochran [2] pág. 50).

$$S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h \doteq P_h Q_h$$

donde $Q_h = 1 - P_h$

De acuerdo con esto obtenemos los siguientes resultados:

Fórmula general (S_h^2 por $P_h Q_h$ en (3.3) y (3.4):

$$n' = \frac{1}{V} \sum \frac{W_h^2 P_h Q_h}{w_h} \quad , \quad n = \frac{n'}{1 + \frac{1}{NV} \sum W_h P_h Q_h}$$

Para costo fijo y varianza mínimas. (S_h^2 por $P_h Q_h$ en (3.10):

$$n = \frac{(C - c_0) \sum (N_h \sqrt{P_h Q_h} / \sqrt{c_h})}{\sum (N_h \sqrt{P_h Q_h} \sqrt{c_h})}$$

Para varianza mínima y costo fijo (S_h^2 por $P_h Q_h$ en (3.11)):

$$n = \frac{\left(\Sigma W_h \sqrt{P_h Q_h} \sqrt{c_h} \right) \Sigma W_h \sqrt{P_h Q_h} / \sqrt{c_h}}{V + \left(\frac{1}{N}\right) \Sigma W_h P_h Q_h}$$

Con afijación óptima de Neyman (S_h^2 por $P_h Q_h$ en (3.13)):

$$n = \frac{\left(\Sigma W_h \sqrt{P_h Q_h} \right)^2}{V + \frac{1}{N} \Sigma W_h P_h Q_h}$$

3.5.- ESTIMACION DE N PARA AFIJACION PROPORCIONAL.

Es frecuente que se prefiera la afijación proporcional, es decir - calcular n_h tal que $n_h/n = N_h/N$, a la afijación desproporcional que -- da algún tipo de afijación óptima. Esto se debe a que la afijación proporcional tiene varias ventajas; una de las principales es que produce estimadores autoponderados lo que reduce considerablemente los problemas de cálculo.

Las fórmulas para n son las siguientes :

Para la estimación de la media $\bar{Y}$ de la población (al sustituir - $w_h = W_h = N_h/N$ en (3.3) y (3.4)):

$$n' = \frac{\Sigma W_h S_h^2}{V}, \quad n = \frac{n'}{1 + \frac{n'}{N}} \qquad (3.14)$$

Para la estimación de la proporción P (si se sustituye $S_h^2 = P_h Q_h$ en (3.14):

$$n' = \frac{\Sigma\, W_h\, P_h\, Q_h}{V}, \qquad n = \frac{n'}{1 + \frac{n'}{N}}$$

C A P I T U L O IV

AFIJACION DE LA MUESTRA MEDIANTE PROGRAMACION NO-LINEAL

4.1.- MUESTREO MULTIVARIADO O DE PROPOSITO MULTIPLE .

En la mayoría de las investigaciones por muestreo es común que se investigue no una sola variable, o característica, sino un grupo de ellas. Por ejemplo, si se tiene una población compuesta por familias, se podría querer investigar varias características: el ingreso familiar, número de personas que componen la familia, nivel de educación, edad, ocupación, -- etc.

En ocasiones la inclusión de varias variables en la investigación se debe a que se quiere aprovechar una misma muestra para varios propósitos, ya que el costo marginal por incluir variables extras puede ser muy bajo. Por ejemplo, en una muestra de predios agrícolas el propósito principal de la investigación podría ser el querer conocer el tamaño promedio de los predios, sin embargo la investigación se puede extender para investigar otras variables que tengan un interés secundario, o para estudiar - otros aspectos, como el cultivo o cultivos que se siembran, tipo de riego, uso de algún implemento agrícola, etc.

En la mayoría de las ocasiones la investigación que se pretende -

realizar no puede concretarse a una sola variable, sino que es necesario analizar varias características para poder realizar un estudio que realmente valga la pena. En el ejemplo anterior, puede ser que resulte de escaso o ningún interés investigar únicamente el tamaño del predio, sin conocer, por ejemplo, el cultivo o cultivos que se siembran, los rendimientos por hectárea o algunas otras variables que se consideren de importancia.

La investigación de varias variables o características complica más el problema de estimar el tamaño de muestra, sobre todo si el diseño de muestra no es irrestricto aleatorio. En un muestreo estratificado tenemos el problema de afijar la muestra, es decir repartir la muestra total en cada uno de los estratos, y un plan de afijación para una variable puede ser totalmente inadecuado para otra variable, aún cuando ambas requieran el mismo tamaño total de muestra.

Los autores que han tratado este problema dan diversas recomendaciones, entre las que destacan las siguientes:

a) Seleccione la variable de más importancia y calcule con respecto a ésta el tamaño de muestra necesario.

b) Calcule el tamaño de muestra para cada variable y de estos tome el máximo.

Estas recomendaciones pueden, en algunos casos, solucionar el problema más o menos satisfactoriamente, pero la mayoría de las veces no dan una solución adecuada. Esto se debe a que, por lo regular, es difícil decidir cuál variable es la más importante, y forzar la selección de una de ellas, puede conducir a escoger una variable que produzca un tamaño de muestra inadecuado para otra variable, la cual, en realidad, también es importante.

El problema de varias variables en un muestreo estratificado es -

tratado por Cochran [2, págs. 120-125], quien ilustra con dos ejemplos - un método sugerido por Yates [12]. Este método, sin embargo, resulta limitado, pues sólamente es aplicable cuando el número de estratos y variables es reducido.

Debido a lo anterior, es deseable encontrar un método que resuelva el problema que se presenta cuando el número de variables y estratos es relativamente grande. En las secciones siguientes se plantea el problema -- del tamaño de muestra y su afijación por estrato, como un problema de programación no-lineal, lo que nos permite tomar en cuenta todas las partes conflictivas que intervienen en la estimación del tamaño de muestra, y encontrar una solución óptima que las concilie, a pesar de que el número de variables sea relativamente grande.

4.2.- LA PROGRAMACION NO-LINEAL .

En términos generales el problema de programación no-lineal posee cuatro componentes fundamentales: un número finito de variables reales, un número finito de restricciones que deben satisfacerse, una función de las variables la cual se quiere optimizar (minimizar o maximizar) y al menos una de las restricciones, o la función a optimizar, es una función no-lineal de las variables.

Lo anterior se puede expresar matemáticamente como sigue:

Encontrar los valores (X_1', X_2', ..., X_n'), *si existen*, de las variables (X_1, X_2, ..., X_n) que satisfacen las restricciones

$$g_i\ (X_1, \ldots, X_n) \leq 0 \qquad i = 1, \ldots, m$$

$$h_j\ (X_1, \ldots, X_n) = 0 \qquad j = 1, \ldots, k$$

y minimizan (o maximizan) la función objetivo

$$f\ (X_1, \ldots, X_n)$$

Donde al menos una de las funciones g_i, h_j o f no puede expresarse como una función lineal de $(X_1, \ldots, X_n)$.

4.3.- FORMULACION DEL PROBLEMA COMO UN MODELO DE PROGRAMACION NO-LINEAL.

Si el índice h denota el estrato y el índice j la variable --- ($h = 1,2,\ldots, L$ y $j = 1,2,\ldots, K$) y además

N = número total de elementos en la población

N_h = número de elementos en el estrato h.

n_h = número de elementos en la muestra del estrato h.

$W_h = \frac{N_h}{N}$ = ponderación por estrato.

Entonces el estimador de la media de la población de la j-ésima variable es:

$$\bar{y}_j = \frac{\sum_{h=1}^{L} N_h \bar{y}_{jh}}{N} = \sum_{h=1}^{L} W_h \bar{y}_{jh} \qquad (4.1)$$

donde $\bar{y}_{jh}$ es media del estrato h de la variable j. La varianza del estimador $\bar{y}_j$ es (Cochran [2], capítulo 5)

$$\text{var}(\bar{y}_j) = \sum_{h=1}^{L} \frac{W_h^2 s_{jh}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h s_{jh}^2}{N} \qquad (4.2)$$

donde s_{jh}^2 es un estimador de la varianza por elemento S_{jh}^2 de la variable j en el estrato h.

En muestreo estratificado los valores de los tamaños de muestra, n_h, pueden fijarse para minimizar la varianza del estimador para un costo específico, o bien minimizar el costo para un valor específico de la varianza del estimador. Cuando se investiga una sola variable (K = 1), y la función de costo es de la forma:

$$C = c_0 + \sum_{h=1}^{L} c_h n_h \qquad (4.3)$$

los tamaños de muestra para cada estrato pueden obtenerse por los métodos estándar, como se hizo en Capítulo 3, en el que se aplicó el método de los Multiplicadores de Lagrange.

Cuando se investigan varias variables ($K > 1$) el problema se complica. Ahora se tiene que minimizar una función de costo, como (4.3), sujeta a restricciones que expresen que las varianzas de los estimadores de las medias de la población deben ser menores o iguales a un valor específico, para cada una de las K variables. Estas restricciones pueden escribirse como :

$$\sum_{h=1}^{L} \frac{W_h^2 \, s_{jh}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h \, s_{jh}^2}{N} \leq V_j \, , \quad j = 1,\ldots,K \qquad (4.4)$$

donde V_j es el límite superior de la varianza de la media de la j-ésima variable. Deberá observarse que en las anteriores restricciones lo único que se desconoce son los valores de n_h ($h = 1,\ldots,L$).

El tamaño de la muestra en el estrato h debe ser no-negativo, o inclusive mayor igual que 2 si se quieren tener cuando menos dos observaciones en cada estrato para poder calcular varianzas, pero además debe ser igual o menor al número total de elementos que contenga el estrato. Estos límites superiores e inferiores pueden expresarse como:

$$2 \leq n_h \leq N_h \, , \qquad h = 1,\ldots,L \qquad (4.5)$$

Por lo tanto el modelo de programación no-lineal para obtener los tamaños de muestra óptimos con respecto a una función de costo, es el siguiente:

Escoger las n_h ($h = 1,\ldots,L$) que minimicen la función de costo (4.3) sujeta a las restricciones no-lineales (4.4) y a los límites (4.5).

4.4.- ALGORITMO USADO PARA ENCONTRAR LA SOLUCION AL PROBLEMA DE PROGRAMACION NO-LINEAL.

El problema de programación no-lineal planteado en la sección anterior requiere de un algorítmo para su solución.

Los resultados que se presentan en las siguientes secciones, son producto de la experimentación con un algorítmo que fué desarrollado e implementado, para ser usado en un equipo UNIVAC 1108, por el Academic Computing Center of The University of Wisconsin [15]

Este algorítmo es una combinación de tres métodos: el método de Gradiente de Rosen, el método de Goldfarb y el método de función de penalidad.

El algorítmo fué adaptado a las condiciones particulares del problema que nos ocupa, y el autor de esta tesis tiene en elaboración un manual para su uso.

4.5.- EJEMPLO CON CUATRO ESTRATOS Y DOS VARIABLES.

Este ejemplo ha sido tomado de Cochran [2, págs. 123-125]. El problema es encontrar el plan de muestreo con costo mínimo, donde las varianzas de los estimadores de las medias de la población son iguales o menores que ciertos valores que se han fijado. La Tabla 4.1 da los datos para el problema, e incluye los tamaños de la población en cada estrato, los valores de las varianzas de las dos variables en los cuatro estratos y el costo unitario de muestrear en cada estrato.

El costo total es igual a $1 + \sum_{h=1}^{4} n_h$ (aquí $c_0 = 1$, $c_1 = 1$, ..., $c_4 = 1$).

T A B L A 4.1

DATOS PARA CUATRO ESTRATOS Y DOS VARIABLES

Estrato h	Población del Estrato N_h	Varianzas de las variables s_{j1}^2	s_{j2}^2	Costo por unidad muestreada C_h
1	400 000	25	1	1
2	300 000	25	4	1
3	200 000	25	16	1
4	100 000	25	64	1

Los límites superiores de las varianzas de los estimadores de las medias de la población para cada variable son:

$$V(\bar{y}_1) \leq .04 \qquad V(\bar{y}_2) \leq .01$$

El problema de programación no-lineal es el siguiente. Escoger n_1, n_2, n_3 y n_4 para minimizar la función

$$(1) + (1)(n_1) + (1)(n_2) + (1)(n_3) + (1)(n_4)$$

sujeta a

$$\frac{(.4^2)(25)}{n_1} + \frac{(.3^2)(25)}{n_2} + \frac{(.2^2)(25)}{n_3} + \frac{(.1^2)(25)}{n_4}$$

$$- \frac{1}{1000\,000}\left[(.4)(25) + (.3)(25) + (.2)(25) + (.1)(25)\right] \leq .04,$$

$$\frac{(.4)^2(1)}{n} + \frac{(.3^2)(4)}{n_2} + \frac{(.2^2)(16)}{n_3} + \frac{(.1^2)(64)}{n_4}$$

$$- \frac{1}{1000\,000}\left[(.4)(1) + (.3)(4) + (.2)(16) + (.1)(64)\right] \leq .01,$$

$$2 \leq n_1 \leq 400.000,$$

$$2 \leq n_2 \leq 300\ 000,$$

$$2 \leq n_3 \leq 200\ 000,$$

$$2 \leq n_4 \leq 100\ 000.$$

La solución a este problema de programación no-lineal se presenta-en la Tabla 4.2. El costo total es $1 + \sum_{h=1}^{4} C_h n_h = 732$. Cochran obtuvo los mismos tamaños de muestra utilizando el método de Yates [14].

T A B L A 4.2

TAMAÑOS OPTIMOS DE MUESTRA Y COSTOS DE MUESTREO POR ESTRATO.

Estrato h	Tamaño óptimo de muestra n_h	Costo $c_h n_h$
1	193	193
2	180	180
3	187	187
4	171	171

4.6.- EJEMPLO CON CATORCE ESTRATOS Y SEIS VARIABLES.

En la sección anterior se presentó un ejemplo hipotético muy sim-- ple con la finalidad de comparar los resultados obtenidos con el método de programación no lineal y el usado por W.G. Cochran. En esta sección se -- presenta un ejemplo que surgió de un problema real de muestreo, en el que- el tamaño de la muestra y su afijación se hizo aplicando la técnica de programación no-lineal, y que, debido al número de variables y estratos, no - hubiera sido posible usar el método utilizado por Cochran.

En esta ocasión se tienen 14 estratos y 6 variables. Los estratos representan clínicas de planeación familiar y las variables, de las cuales se quiere estimar su media, son las siguientes:

1. Edad de las pacientes

2. Ingreso familiar

3. Número de hijos vivos

4. Número de embarazos

5.- Número de abortos

6. Número de años de educación.

La Tabla 4.3 muestra los datos que corresponden a cada una de estas variables y estratos.

Los resultados obtenidos por programación no-lineal se comparan en las Tablas 4.5 y 4.6 con los obtenidos por otros dos posibles métodos. La Tabla 4.4 da los tamaños de muestra basados en la afijación óptima de Ney-

T A B L A 4.3

DATOS PARA EL EJEMPLO DE UN DISEÑO ESTRATIFICADO CON SEIS CARACTERISTICAS

Estrato h	Tamaño de la población en el estrato h N_h	Costo (\$) C_h	Desviaciones estimadas de las características s_{h1}	s_{h2}	s_{h3}	s_{h4}	s_{h5}	s_{h6}
1	2902	18.3	22.9	517	2.4	2.7	.67	2.5
2	7015	18.5	11.3	592	5.8	2.5	.62	2.1
3	9542	23.6	22.9	522	5.6	3.4	.12	2.6
4	4947	22.5	10.3	893	3.3	3.2	.22	5.0
5	3005	17.1	9.6	609	3.2	6.2	.15	4.8
6	6132	20.5	8.3	542	2.6	6.7	.13	4.9
7	5717	21.6	8.2	872	2.1	3.1	.19	4.8
8	12032	18.7	11.1	812	3.4	6.8	.16	2.1
9	1372	23.9	8.2	899	3.1	6.7	.09	2.0
10	1214	15.2	23.5	501	2.2	2.1	.63	2.2
11	13931	32.3	21.4	602	5.2	2.9	.19	3.0
12	3868	17.3	23.7	583	3.0	2.6	.64	1.9
13	8912	25.6	10.3	612	5.7	2.0	.65	2.9
14	11358	29.0	9.7	884	3.3	2.6	.05	4.7
Varianza especificada			0.51	708	0.022	0.02	.0001	0.0144
Coeficiente de variación especificado[(a)]			3	3	3	3	3	4

T A B L A 4.4

AFIJACION OPTIMA DE NEYMAN PARA CADA UNA DE LAS CARACTERISTICAS

	Característica					
Estrato	1	2	3	4	5	6
1	23	18	16	17	63	20
2	27	49	93	39	139	41
3	66	52	108	63	32	61
4	16	47	34	32	32	63
5	10	22	23	43	15	42
6	17	37	34	86	24	80
7	15	59	25	36	32	71
8	45	114	93	179	71	70
9	3	13	9	18	3	7
10	11	8	7	6	27	8
11	77	75	125	67	64	88
12	32	27	27	23	82	21
13	27	54	98	33	158	61
14	30	94	68	52	15	119
TOTAL	399	664	760	694	747	752

T A B L A 4.5

TAMAÑOS DE MUESTRA CON TRES METODOS DE AFIJACION

	Método		
Estrato	Neyman, para la característica que requiere el tamaño de muestra más grande.	Programación no-lineal	Estrato máximo en la afijación de Neyman.
1	25	46	63
2	23	105	139
3	9	68	108
4	93	51	63
5	34	36	43
6	27	68	86
7	68	54	71
8	7	104	179
9	34	9	27
10	16	22	27
11	108	92	125
12	98	61	82
13	125	122	158
14	93	86	119
Tamaño total de muestra	760	924	1281
Costo ($)	*18,060*	*20,852*	*28,864*

T A B L A 4.6

COEFICIENTES DE VARIACION (%) DE LOS ESTIMADORES DE LAS MEDIAS DE CADA UNA DE LAS CARACTERISTICAS PARA LOS TRES METODOS DE AFIJACION

		Método		
Característica	Máximo especificado	Neyman, para la característica que requiere el tamaño de muestra más grande.	Programación no-lineal	Estrato máximo en la afijación de Neyman.
1	3.0	2.4	2.3	1.9
2	3.0	3.1*	2.7	2.3
3	3.0	3.0	3.0	2.5
4	3.0	3.4*	3.0	2.5
5	3.0	3.8*	3.0	2.6
6	4.0	4.7*	4.0	3.4

* Excede el máximo especificado

man aplicada a cada una de las características por separado. La afijación óptima de Neyman requerida para la muestra total más grande es la que está basada en la característica 3; W. G. Cochran [2] señala que tal afijación puede satisfacer la variación especificada para las demás variables, esto no sucede aquí como puede apreciarse en la Tabla 4.6. En la última columna de la Tabla 4.5, aparece el tamaño de muestra y la afijación que corresponde al máximo tamaño de muestra por estrato, según las afijaciones individuales de Neyman de la Tabla 4.4. Esta afijación de "estrato máximo" -- produce varianzas más pequeñas que las especificadas, como puede apreciarse en la Tabla 4.6, sin embargo el tamaño de muestra requerido es demasiado grande e innecesario, ya que la afijación por el método de programación no-lineal produce un tamaño menor de muestra sin exceder los límites especificados de las varianzas de los estimadores, o de sus correspondientes - coeficientes de variación.

4.7.- EJEMPLO CON TRECE ESTRATOS Y SIETE VARIABLES.

Este último ejemplo está elaborado con datos que provienen del Servicio de Estadísticas Agropecuarias del Departamento de Agricultura de los Estados Unidos.

La presentación de un ejemplo más puede parecer ociosa, sin embargo los resultados que se obtienen con estos datos hacen destacar el beneficio que se puede obtener al afijar la muestra con métodos de programación-no-lineal.

Los datos, Tabla 4.7, se refieren a los inventarios de siete artículos agrícolas en una población dividida en trece estratos.

Como en el ejemplo de la sección anterior, en las Tablas 4.9 y 4.10 se comparan los resultados obtenidos por programación no-lineal con otros dos métodos. La Tabla 4.8 da los tamaños de muestra basados en la afijación Optima de Neyman, aplicada a cada una de las características por separado.

En la Tabla 4.10 pueden observarse los grandes coeficientes de variación que se obtienen cuando se usa la afijación Optima de Neyman, para la característica que requiere el tamaño de muestra más grande. Si se escogiera esta afijación, la variable número 3 se estimaría con un coeficiente de variación del 30.5%, en lugar del 10% que se ha fijado como máximo.

La afijación de estrato máximo, columna 3, proporciona coeficientes de variación un poco menores a los máximos especificados. Sin embargo, el tamaño total de muestra correspondiente, y por lo tanto el costo, son mucho mayores a los encontrados por programación no-lineal, y, desde luego, este último método proporciona coeficientes de variación menores o iguales a los límites que se han fijado.

T A B L A 4.7

DATOS PARA EL EJEMPLO DE TRECE ESTRATOS Y SIETE CARACTERISTICAS

Estrato	Tamaño de la población en el estrato h.	Costo ($)	Desviaciones estándar de las características						
h	N_h	C_h	S_{h1}	S_{h2}	S_{h3}	S_{h4}	S_{h5}	S_{h6}	S_{h7}
1	4714	21.2	1.0	311	27	161	551	30	350
2	5718	20.6	4.5	70	1	208	27	9	331
3	4686	18.3	1.2	135	80	126	152	13	65
4	6134	20.8	1.2	265	266	86	115	99	50
5	9912	23.6	8.6	116	78	35	79	55	24
6	28044	19.7	2.0	74	65	44	34	49	45
7	24642	16.2	2.2	75	1	5	1	5	2
8	11328	20.5	4.5	98	1	13	1	19	8
9	1144	35.3	5.1	844	2015	1386	1	4	372
10	4948	23.7	2.5	321	2507	81	30	91	123
11	14932	18.6	1.7	98	22	43	1	64	86
12	1378	18.1	1.4	88	3	293	22	7	488
13	7016	16.3	4.9	190	6	88	2	27	71
Varianza especificada			.009	29.6	24.7	4.7	2.8	1.5	8.3
Coeficiente de variación especificado (%)			5	6	10	5	6	8	6

T A B L A 4.8

AFIJACION OPTIMA DE NEYMAN PARA CADA UNA DE LAS CARACTERISTICAS.

	Característica						
Estrato	1	2	3	4	5	6	7
1	12	49	6	88	349	28	115
2	67	14	1	141	21	10	134
3	15	23	20	74	103	13	23
4	19	55	83	62	96	119	22
5	206	37	37	38	100	101	16
6	148	72	95	149	133	278	92
7	158	71	1	16	4	27	4
8	132	38	1	17	2	43	6
9	12	25	90	143	1	1	23
10	29	51	590	44	19	83	40
11	69	53	18	80	2	199	96
12	5	4	1	51	4	2	51
13	100	51	2	82	2	42	40
TOTAL	972	543	945	985	836	946	662

T A B L A 4.9

TAMAÑOS DE MUESTRA CON TRES METODOS DE AFIJACION.

	Método		
Estrato	Neyman, para la característica que requiere el tamaño de muestra más grande.	Programación no-lineal	Estrato máximo en la afijación de Neyman.
1	88	253	349
2	141	82	141
3	74	85	103
4	62	114	119
5	38	170	206
6	149	216	278
7	16	111	158
8	17	97	132
9	143	102	143
10	44	476	590
11	80	114	199
12	51	24	51
13	82	79	100
Tamaño total de muestra	985	1923	2569
Costo ($)	21692	41758	55553

T A B L A 4.10

COEFICIENTES DE VARIACION (%) DE LOS ESTIMADORES DE CADA UNA DE LAS CARACTERISTICAS PARA LOS TRES METODOS DE AFIJACION.

		Método		
Característica	Máximo Especificado	Neyman, para la característica que requiere el tamaño de muestra más grande.	Programación no-lineal	Estrato máximo en la afijación de Neyman
1	5	10.2*	5.0	4.3
2	6	6.7*	3.6	3.1
3	10	30.5*	10.0	8.8
4	5	5.0	5.0	4.1
5	6	9.6*	6.0	5.2
6	8	11.3*	7.9	6.8
7	6	5.6	5.3	4.1

* Excede el máximo especificado

4.8.- TIEMPO DE COMPUTO .

A continuación se dan algunos datos sobre el tiempo de cómputo requerido por el algorítmo de programación no-lineal para obtener la solución. Esto es importante, ya que la especificación de las varianzas V_j pueden, en la práctica, ser algo imprecisas; lo que puede requerir soluciones alternativas para diferentes valores de V_j, pero el cómputo puede ser demasiado caro para que puedan considerarse un buen número de ellas.

La Tabla 4.11 da una indicación del monto de tiempo requerido por un equipo UNIVAC 1106 para resolver tres problemas; los dos primeros se han descrito en este trabajo.

T A B L A 4.11

TIEMPO DE COMPUTO REQUERIDO PARA OBTENER LA AFIJACION POR-PROGRAMACION NO-LINEAL

Número de Estratos.	Número de características	Tiempo requerido para su solución (segundos)
4	2	19.5
14	6	37.2
30	8	65.3

C O N C L U S I O N E S

La determinación del tamaño de muestra es un problema ineludible - en toda investigación por muestreo. Existen varios factores que, directa- o indirectamente, afectan el número de elementos que deben constituir la - muestra; estos factores están interrelacionados y su importancia relativa- varía en cada caso.

El cálculo del tamaño de muestra plantea problemas mayores a medida que el diseño de muestreo se complica.

En particular, la determinación del tamaño de muestra cuando se investigan varias características, y se usa un diseño estratificado, plantea un problema difícil de solucionar. En este trabajo se ha formulado este - problema como un modelo de programación no-lineal, y con la aplicación de- un algorítmo desarrollado por la Universidad de Wisconsin, se presentan varios ejemplos en los cuales se comparan los resultados obtenidos con otros dos posibles métodos. Se concluye que el beneficio de usar programación-- no-lineal para resolver este problema puede llegar a ser significativamente grande.

B I B L I O G R A F I A

[1] Bracken, J. and McCormick, G.P. (1968). Selected Applications of Nonlinear Programming. Wiley.

[2] Cochran, W.G. (1963). Sampling Techniques, 2nd. Ed. Wiley.

[3] Delanius, T. and Hodges, J.L., Jr. (1959). "Minimun variance stratification". Jour. Amer. Stat. Assoc., 54, 88-101.

[4] Garza, T. y Coronel, J.A. (1970). "Un método para la determinación del tamaño de muestra en encuestas sobre poblaciones finitas". Demografía y Economía, El Colegio de México, 10, 121-128.

[5] Hartley, H.O. (1965). "Multiple purpose optimum allocation in stratified sampling". Proc. Amer. Statist. Ass., Social Statistics Section, 258-261.

[6] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vols. I and II. New York: John Wiley and Sons.

[7] Jagannathan, R. (1965). "The programming aproach in multiple character studies". Econometrica, 33, 263-237.

[8] Kish, L. (1965). Survey Sampling. Wiley.

[9] Kish, L. (1969). "Design and estimation for subclasses, comparisons, and analytical statistics", New Developments in Survey Sampling (a symposium), University of North Carolina. Wiley - Interscience.

[10] Kokan, A.R. (1963), "Optimum allocation in multivariate Surveys". J. Roy. Statist. Soc., Ser. A, 126, 557-565.

[11] Kokan, A.R. and. Khan, S. (1967). "Optimum allocation in multivariate surveys. an analytical solution". J.R. Statist. Soc. B, 29, 115-175.

[12] Rosen, J.B. (1960). "The Gradient Projection Method for Nonlinear Programming, Part. I: Linear Constraints", J. Soc. Ind.- Appl. Math., 9, 181-217.

[13] Rosen, J.B. (1961). "The Gradient Projection Method for Nonlinear programming, Part. II: Nonlinear Constraints", J. Soc. - Ind. Appl. Math., 9, 514-532.

[14] Yates, F. (1960). Sampling Methods for Censuses and Surveys. Charles Griffin and Co. London. Third Edition.

[15] GPM/GPMNLC. Extended Gradient Projection Method, Nonlinear - Programming Subroutines. Reference Manual for the 1108. Academic Computing Center. The University of Wisconsin - Madison.