UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Facultad de Ciencias

CONSIDERACIONES SOBRE LA ESTRATIFICACION EN ENCUESTAS POR MUESTREO

T E S I S

Que para obtener el titulo de

A C T U A R I O

P r e s e n t a

Rebeca San Juan Téllez.

México, D.F. 1982.

I

CONSIDERACIONES SOBRE LA ESTRATIFICACION EN ENCUESTAS POR MUESTREO

INTRODUCCION

La teoría y aplicación de muestreo es, quizá, una de las ramas más antiguas de la teoría estadística; casi cualquier aspecto en fenómenos naturales, humanos, y en otras actividades, está sujeta a medidas en términos estadísticos; las cuales, en ocasiones, se buscan interpretar sistemáticamente.

En épocas lejanas el muestreo ya era usado(Armitage); sin embargo, es difícil encontrar datos que revelen quién fué el primero que lo utilizó. De cualquier forma, sabemos que los primeros hombres de ciencia bajo un procedimiento de selección, obtenían conclusiones de sus estudios sobre las leyes de la Naturaleza al hacer observaciones del medio ambiente(Flores).

Entre los primeros investigadores que se sabe usaron técnicas de muestreo se encuentran, - por ejemplo, Halley(1693) quién trató de estudiar la mortalidad de la humanidad, Sir Frederick Moston Eden (1800) estimó la población de Inglaterra, confirmando

sus datos en el censo británico en 1801, Sir Jones Lawes estimó las cosechas de trigo de 1852 a 1879 en Inglaterra y Gales; Lavoisier usando muestras quiso saber la superficie de tierras cultivadas, número de caballos, ganado, carneros, etc, que había en Francia; Le Play hizo investigaciones sanitarias y sociales; - Broth investigó vida y trabajo en Londres; Rowentree - investigó la pobreza en York.

En este siglo, durante la tercera década hubo una gran demanda en cuanto al uso de los métodos de muestreo para investigaciones sociales y económicas en los Estados Unidos, En la oficina de Economía Agrícola(IOWA State University), se estimaron características de ranchos, de producción de ganado y cosechas; se usaron además censos muestrales en Polonia, Bulgaría y Suecia, en India se desarrollaron nuevos métodos de muestreo para estimar la producción de cosechas.

Al mismo tiempo los científicos sociales experimentaron con métodos muestrales y con técnicas de pregunta-respuesta y los investigadores de - Rothamested Agricultural Experimental Station en Inglaterra desarrollaron nuevos métodos y avanzaron en la - teoría de sub-muestreo y muestreo doble(Jessen).

El uso del muestreo -como hemos visto- fué consecuencia de una necesidad de los estadísticos al enfrentarse a problemas de población, industria, - finanzas, agricultura, sanidad pública, etc, de hecho, el progreso de la teoría y métodos para investigación estadística del muestreo se ha llevado a cabo en los últimos 75 años.

Entre los precursores en este siglo se encuentran los ingleses Fisher, Cochran, Hansen, - Hurwitz y Madow, Jerszy Neyman; los hindues Hubback , Lahiri y Mahalanobis. En la actualidad se distinguen - significativamente Leslie Kish, Martin Frankel, -- Graham Kalton, Des Raj, Sukhatme, T.M.Smith, T.C.Koop, Cumberland, Jessen.

Los objetivos de la investigación se suelen evaluar -en el muestreo- en términos de precisión; una medida de precisión del estimador es una medida de proximidad entre el conjunto de posibles estimadores muestrales y el promedio de una serie de medidas llevadas a cabo bajo condiciones similares - Cochran(1957).

También podemos definir -y es ilustrativo- la precisión como el inverso de la varianza de

las estimaciones de la encuesta.

La exactitud, por su parte, es el inverso del error total que incluye tanto el sesgo como la varianza; si hay sesgos importantes, especialmente errores de no muestreo y son además "distinguibles", la exactitud es una mejor medida de los objetivos de la encuesta que la precisión.

El diseño de la muestra busca tener precisión máxima(varianza mínima). Además se debe incluir el criterio del costo; una muestra se dice que es "económica" o de costo mínimo cuando la precisión por unidad de costo es alta, o el costo por unidad de varianza es bajo.

Usando los conceptos antes mencionados, se intentará mostrar, de manera general en este trabajo, que la asignación del tamaño de muestra(en particular, del muestreo estratificado) debe tomar como punto de partida los tipos de encuesta, a saber, si es analítica o descriptiva.

En el capítulo II se mencionan las diferencias en los criterios de asignación según el tipo de encuesta.

En el capítulo III, se aclara suscintamente lo que es el muestreo y las encuestas, además - se exhiben los tipos de encuesta, los fundamentos, ventajas y métodos de estratificación en el muestreo.

En los capítulos IV y V se describen - algunos métodos a seguir para la asignación de muestra según el tipo de encuesta con que se esté trabajando.

Finalmente, el capítulo VI presenta un caso particular del capítulo V, y se refiere al análisis de dominios de estudio con multiplicidad de objetivos.

II. DIFERENCIAS EN LOS CRITERIOS DE ASIGNACION SI LA ENCUESTA POR MUESTREO ES ANALITICA O DESCRIPTIVA.

El objetivo de la investigación es descubrir respuestas a determinadas interrogantes a través de la aplicación de procedimientos científicos. La investigación procede a la concepción de un tema- y se alimenta por los estudios realizados mediante la acumulación de datos hasta la elaboración de un informe y la aplicación de sus resultados.

La inferencia se usa siempre que los - datos de la investigación están basados en muestras; - en este caso el investigador tiene la tarea de esti - mar las características de la población, de la muestra y los errores muestrales y no muestrales.

Los problemas de inferencia surgen en diferentes contextos, es decir: i) al trabajar con una medida de correlación entre dos variables de una - investigación y ii) al buscar una explicación por medio de métodos analíticos(tal como el análisis de correlación). El uso de técnicas estadísticas "complica

das" en las investigaciones se debe al deseo de establecer e interpretar relaciones multivariadas(Moser - y Kalton).

El objetivo de la investigación determina la unidad de muestreo a seleccionar; cualquiera que sea la unidad de muestra es importante poseer una base para identificar la población total y un método adecuado para seleccionar dichas unidades a partir de tal población.

Las dos consideraciones básicas en la selección de una muestra son:

i) La presencia o ausencia del mecanismo de selección.

ii) La objetividad y/o subjetividad por parte del investigador, es decir, es objetivo cuando las características a medir son claras y no hay ambigüedad para la selección; es subjetivo cuando el investigador se sirve de su juicio para seleccionar lo que él considere una "buena muestra".

Al considerar los factores anteriores, a dos niveles tenemos los siguientes tipos de muestras:

Enfoque del Investigador	Procedimiento de Selección	
	Probabilístico	No Probabilístico
Objetivo	Muestras Aleatorias	Muestras con selección deliberada.
Subjetivo	Muestras cuasi-aleatorias	Muestras "de juicio"

Estos cuatro tipos de muestra son muy-usados en la práctica. Los investigadores prefieren - las muestras aleatorias debido a la teoría con que se cuenta para trabajar y así entender el comportamiento de tales muestras(Jessen).

Cada diseño de muestreo puede estar inscrito en un diseño de encuesta. La teoría de encuestas muestrales se ha desarrollado, principalmente, para proporcionar estadísticas descriptivas, sobre todo de medias y totales. Los diseños experimentales se han usado fundamentalmente en la búsqueda analítica de variables explicativas y para la medición de relaciones, por lo tanto, son las encuestas y no los experimentos los que sirven frecuentemente de instrumento de análi

sis en muchos campos; entre ellos, las ciencias sociales. En muchas ocasiones ni los experimentos ni las encuestas son suficientes para el análisis, por lo que se recurre a otro tipo de investigación.

El objetivo del estudio descriptivo es aumentar la "familiaridad" con el análisis de la investigación, toma consideraciones de economía y error; dicho análisis se hace de manera general, esto supone procedimientos de estimación conocidos, y a partir de los resultados de la muestra y de la frecuencia de alguna característica en la población obtener información de los "grandes grupos" en estudio.

Los procedimientos a usar en el estudio descriptivo deben ser cuidadosamente planificados. Si tenemos en cuenta que el objetivo es obtener información completa y exacta, el proyecto de investigación debe tomar las medidas contra los errores de sesgo.

En cambio en los estudios analíticos se desea tener una medida del "grado y dirección" de las relaciones entre dos o más variables dentro de los diferentes subgrupos de la población, para verificar hipótesis relacionadas con el objeto de estudio; entre los métodos más valiosos usados en este tipo de estu-

dio se pueden citar los de: regresión, correlación, análisis discriminante, análisis de varianza y algunos métodos relacionados con diseños de experimentos(Kish y Purcell), sin embargo, la teoría carece de los elementos suficientes para estudios analíticos y profundos de muestras complejas, ,por lo que los investigadores realizan el análisis con las fórmulas hasta ahora conocidas y a la vez se plantean nuevos problemas en la teoría estadística.

La estratificación incrementa la varianza entre las medias de los estratos, pero cuando se miden varias características de la población, el análisis se vuelve complejo(Dominios de estudio).

Si dos características están perfectamente correlacionadas, ya sea positiva o negativamente, cualquier modo de estratificación, mejora las estimaciones; por el contrario, si las dos características no están correlacionadas pueden o no mejorar las estimaciones(Evans).

En los estudios descriptivos las correlaciones entre variables de estratificación y variables de estimación son notorias; estas relaciones tienen importancia para estimaciones de variables en futuras investigaciones(Hoss,Sethi).

III. LA ESTRATIFICACION EN LA POBLACION OBJETO

III.I FUNDAMENTOS Y VENTAJAS

Es común, tanto en los problemas cotidianos como en la investigación científica, el empleo de muestras para inferir alguna o algunas propiedades del universo de donde se obtienen.

Uno de los propósitos más importantes del muestreo es desarrollar métodos de selección de muestras y métodos de estimación que proporcionen estimadores de costo mínimo y suficientemente precisos en relación al propósito buscado.

Tal y como se ha referido anteriormente, los diseños de encuesta y los de muestreo están íntimamente vinculados; las encuestas en general, y por muestreo en particular, se clasifican en descriptivas y analíticas. En una encuesta descriptiva el objetivo es simplemente obtener información acerca de grandes grupos o segmentos de la población objeto; por ejem -

plo para conocer: la producción total de trigo en -- cierta entidad, cuáles son los ingresos medios de las personas que trabajan, niveles de escolaridad en los habitántes de una ciudad o país, incidencia o preva - lencia de drogadicción en una ciudad o país, y algu - nos otros aspectos demográficos y sociales.

En este tipo de encuestas al contestar como se distribuyen las características en la pobla - ción, confirma sus carácter descriptivo exploratorio ; es decir, la investigación exploratoria es necesaria - para obtener la formulación de hipótesis relevantes ,- para una investigación más definitiva, aunque en tér - minos generales se discute el estudio exploratorio co mo una entidad en sí misma, es adecuado considerarlo como un proceso continuo de investigación(Selltiz).

Por otra parte, en una encuesta analí - tica, se hacen comparaciones entre subgrupos diferen- tes de la población para descubrir si la diferencia entre ellos nos permiten formar o verificar hipótesis (Cochran,1976). La base esencial de las encuestas ana líticas es, entonces, el planteamiento explícito de - hipótesis.

Un esquema sobre las etapas en una encuesta por muestreo es:

a)"Objetivos de la encuesta,
b) Población bajo muestreo,
c) Marco muestral,
d) Unidad muestral,
e) Selección muestral,
f) Información a seleccionar,
g) Métodos de observación(Medición),
h) Cuestionario,
i) Encuesta piloto,
j) Organización de trabajo de campo,
k) Resumen y analísis de los datos,
l) Información ganada para futuras encuestas(Des Raj)".

Los objetivos de la encuesta deben determinar el diseño de la muestra; pero puede ocurrir, que el proceso sea reversible, puesto que los problemas del diseño de muestra puede influir e incluso cambiar los objetivos de la encuesta(Kish,1962). El caso más simple, se dá cuando por la cobertura del marco - muestral se tenga necesidad de redefinir la población de encuesta.

En el muestreo estratificado es posible subdividir una población heterogénea en subpoblaciones, cada una de las cuales se pretende sea internamente - homogénea, dichas subpoblaciones son llamadas "estra- tos".

La estratificación es un procedimiento común en la teoría de muestreo, nos permite seleccio- nar un segmento o grupo estratificado con probabili - dad igual a 1, de aquí que las diferentes partes rele vantes de la población, tanto por razones estadísti - cas(por ejemplo, los estratos como dominios de estu - dio), como de índole administrativa o de control que pudiesen integrar estratos; dichos subconjuntos están formados según ciertas características o variables - que son consideradas importantes.

Las razones principales por las que se recurre a la estratificación son:

1) Se utiliza para disminuir la varian za total de los estimadores y, consecuentemente, de - la muestra. En el esquema de estratificación propor - cional, el tamaño de muestra que se selecciona de ca da estrato, es proporcional al tamaño de la población.

Así, la varianza total se disminuye — mientras aumente el grado de heterogeneidad entre cada estrato. Es decir, mientras los estimadores obtenidos para cada estrato sean producto de subpoblaciones que discrepan entre sí, la varianza del estimador global se reducirá.

2) Cuando se requieren estimadores por separado para cada subdivisión, tales como: subdivisio geográficas, o cuando interesa conocer alguna característica de los hogares de una ciudad, por ejemplo gastos en alimentos, ropa etc; ingresos tipos de casa-habitación, años de escolaridad del padre, número de - hijos, etc. Se sabe que estas características depen - den del nivel socioeconómico de las familias, por lo tanto conviene hacer estratos considerando áreas de - la ciudad con niveles socioeconómicos semejantes. Así las colonias se pueden clasificar apriori en relación al nivel socioeconómico como: Muy alto, alto, medio, - medio-bajo, y bajo, formando así cinco estratos. En - general, podemos decir que son subclases que configuran dominios de estudio.

3) Cuando se requieren estimadores de - alta eficiencia por unidad de costo; se ha recomenda-

do utilizar como base para primeras etapas en estratificaciones.

4) Cuando se tienen que usar diferentes procedimientos de muestreo para diferentes subuniversos formados según la naturaleza de la información disponible para la selección de muestra(Ranjar)".

La encuesta se planea para cada estrato por separado. Otra razón poderosa para formar estratos es la disponibilidad de marcos. Así, si para una parte de la población tenemos un buen marco, lo usamos para el muestreo de esa parte y la o las otras partes de la población las muestreamos usando otros marcos más imprecisos y, posiblemente otros esquemas(diseños) de muestra.

Otra razón más para construir estratos puede ser el costo de localizar y levantar la información de las unidades(en una encuesta de predios agrícolas cuyo acceso es difícil, esa región puede constituir un estrato).

Se pueden usar diferentes formas de muestreo en los diferentes estratos, sin embargo, en este trabajo considerése el caso general con la siguiente notación.

NOTACION

Sea	h	h-ésimo estrato.
	i	i-ésima unidad dentro del estrato.
	hi	i-ésima unidad dentro del h-ésimo estrato.
	N_h	Número de unidades elementales o de última etapa en el estrato(el subíndice h indica un estrato en particular).
	n_h	Número de unidades en la muestra.
	y_{hi}	Valor obtenido para la i-ésima unidad(con respecto a la característica en estudio).
	$W_h = N_h / N$	Ponderación del estrato.
	$f_h = n_h / N_h$	Fracción de muestreo en el estrato.
	$\bar{Y}_h = \sum Y_{hi}/N_h$	Media verdadera.
	$\bar{y}_h = \sum y_{hi}/n_h$	Media muestral.
	$S_h^2 = \sum (y_{hi}-\bar{Y}_h)^2 / N_{h-1}$	Varianza verdadera.

III.2 PROPIEDADES DE LOS ESTIMADORES

Para la media de la población por unidad el estimador para el muestreo estratificado es:

$$\bar{y}_e = \frac{\sum N_h \bar{y}_h}{N}$$

donde $N = N_1+N_2+\ldots\ldots+N_L$, i.e. $N= \sum N_h$

La media de la muestra es:

$$\bar{y} = \frac{\sum n_h \bar{y}_h}{n}$$

La diferencia entre esos dos estimadores es que $\bar{y}_e$, el estimador obtenido de estimadores individuales tiene las ponderaciones correctas para obtener un estimador insesgado; por otro lado, $\bar{y}$ & $\bar{y}_e$ coinciden siempre y cuando la fracción de muestreo sea la misma para todos los estratos, es decir:

$$n_h / n = N_h / N \quad \text{ó} \quad n_h / N_h = n / N$$

Debe hacerse notar que:

$$E[\bar{y}] = \sum \frac{n_h}{n} \bar{Y}_h \neq \bar{y}_e = \sum \frac{N_h}{N} \bar{Y}_h \text{ si } \exists i \ni \frac{n_i}{n} \neq \frac{N_i}{N}$$

TEOREMA 1 Si en todos y cada uno de los estratos el estimador $\bar{y}_h$ es insesgado, entonces $\bar{y}_e$ es un estimador insesgado de la media de la población $\bar{Y}$.

DEMOSTRACION

$$E(\bar{y}_e) = E\left(\frac{\sum N_h \bar{y}_h}{N}\right) = \frac{\sum N_h \bar{Y}_h}{N}$$

ya que los estimadores son insesgados en los estratos individuales.

La media de la población $\bar{Y}$ se puede escribir:

$$\bar{Y} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}}{N} = \frac{\sum N_h \bar{Y}_h}{N}$$

COROLARIO Como $\bar{y}_h$ es un estimador insesgado de $\bar{Y}_h$ en el muestreo simple aleatorio dentro de un estrato $\bar{y}_e$ es un estimador insesgado de $\bar{Y}$ en el muestreo estratificado aleatorio.

TEOREMA 2 En el muestreo estratificado, la varianza de $\bar{y}_e$, como un estimador de la población $\bar{Y}$, es:

$$V(\bar{y}_e) = \frac{\sum N_h^2 V(\bar{y}_h)}{N^2} = \sum W_h^2 V(\bar{y}_e)$$

donde *$V(\bar{y}_h) = E(\bar{y}_h - \bar{Y}_h)^2$, existen dos restricciones en el teorema anterior.

1) y_h debe ser un estimador insesgado de $\bar{Y}_h$ (sólo para lo señalado en *).

2) las muestras deben obtenerse independientemente en los estratos en general.

DEMOSTRACION

$$\bar{y}_e - \bar{Y} = \frac{\sum N_h \bar{y}_h}{N} - \frac{\sum N_h \bar{Y}_h}{N}$$

$$= \frac{\sum N_h (\bar{y}_e - \bar{Y}_h)}{N}$$

Como el error $(\bar{y}_e - \bar{Y})$ en el estimador es expresado como media ponderada de los errores de la estimación que han sido hechos dentro de los estratos individuales, por lo tanto:

$$(\bar{y}_e - \bar{Y})^2 = \frac{\sum N_h^2 (\bar{y}_h - \bar{Y}_h)^2}{N^2} + \frac{2 \sum N_h N_j (\bar{y}_h - \bar{Y}_h)(\bar{y}_j - \bar{Y}_j)}{N^2}$$

donde el segundo término cubre todas las parejas de estratos.

Al promediar sobre todas las muestras en cualquier término de doble producto, manteniendo la muestra en el estrato h fija y promediando sobre todas las muestras en el estrato j; Como el muestreo es independiente en los dos estratos las posibles muestras en el estrato j serán las mismas y tendrán las mismas probabilidades cualquiera que haya sido la muestra obtenida en el estrato h.

Como $\bar{y}_j$ es insesgada, el promedio $(\bar{y}_j - \bar{Y}_j)$ es cero, por lo tanto los términos del doble producto desaparecen y los términos cuadráticos nos dan:

$$V(\bar{y}_e) = \frac{\sum N_h^2 E(\bar{y}_h - \bar{Y}_h)^2}{N^2} = \frac{\sum N_h^2 V(\bar{y}_h)}{N^2}$$

Cabe señalar que la varianza de $\bar{y}_e$ sólo depende de los estimadores de las medias de los estratos individuales $\bar{Y}_h$

TEOREMA 3 En el muestreo estratificado aleatorio, la varianza del estimador $\bar{y}_e$ es:

$$V(\bar{y}_e) = \frac{1}{N^2} \sum N_h (N_h - n_h) \frac{S_h^2}{n_h} = \sum W_h^2 \frac{S_h^2}{n_h} (1 - f_h) \qquad (a)$$

DEMOSTRACION

Para un estrato particular tenemos que:

$$V(\bar{y}_h) = \frac{S_h^2}{n_h} \frac{N_h - n_h}{N_h}$$

Sustituyendo en el teorema anterior(ya que $\bar{y}_h$ es estimador insesgado de $\bar{Y}_h$) tenemos:

$$V(\bar{y}_e) = \frac{1}{N^2} \sum N_h^2 V(\bar{y}_h)$$

$$= \frac{1}{N^2} \sum N_h^2 \left[\frac{S_h^2}{n_h} \frac{N_h - n_h}{N_h} \right]$$

$$= \frac{1}{N^2} \sum N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

$$= \sum W_h^2 \frac{S_h^2}{n_h} (1 - f_h)$$

Para casos particulares tenemos los siguientes corolarios:

COROLARIO Con asignación proporcional, sustituimos - en (a): $n_h = nN_h / N$, obteniendo:

$$V(\bar{y}_e) = \sum \frac{N_h}{N} \frac{S_h^2}{n} \left(\frac{N-n}{N}\right)$$

$$= \frac{1-f}{n} \sum W_h S_h^2$$

COROLARIO Si el muestreo es proporcional y las varian zas en todos los estratos tienen el mismo - valor, S_w^2 se obtiene:

$$V(\bar{y}_e) = \frac{S_w^2}{n} \left(\frac{N-n}{N}\right)$$

TEOREMA 4 Si $\hat{Y}_e = N\bar{y}_e$ es el estimador del total - de la población Y entonces:

$$V(\bar{y}_e) = \sum N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

III.3 METODOS DE ESTRATIFICACION

Para obtener una estratificación efi - ciente, se debe

1) Procurar que las unidades dentro de los estratos sean lo más homogéneas posibles y

2) Que las medias entre los estratos - difieran tanto como sea posible.

Dos de los métodos usados con más fre- cuencias son:

a) Estratificación geográfica y

b) Usar la información de alguna variable correlacionada con la variable en estudio, como - base para la estratificación. Sin embargo, la práctica ha demostrado que en la estratificación geográfica el beneficio que se obtiene es mínimo; en cambio, al usar alguna característica correlacionada con algún elemento de interés aumenta considerablemente la pre - cisión.

La contribución de la estratificación - a la precisión del diseño muestral radica principal - mente en la relación entre las variables de estratificación y las variables de estudio.

Para la formación de estratos es posible establecer algunos principios que se pueden usar como pautas o guías:

1) Casi todos los marcos de referencia poseen cierta "estratificación natural", es decir, dependen del orden en que las unidades aparecen en el marco; dicho orden proporciona excelente estratificación geográfica y con frecuencia estratificación por tipo y tamaño.

La ventaja en este caso es que, si el marco no está preparado, se pueden formar los estratos de acuerdo a los objetivos de la investigación a bajo costo; considerando principalmente, la configuración - del marco como etapa previa al diseño. Es cierto que - en unidades geográficas comunicadas y no extensas relativamente, sucesivas etapas de estratificación pueden inducir a tener costos más reducidos, pero, por ejemplo, en México, esta no es una regla general y dependerá de las actividades de logística inherentes al diseño de la encuesta.

2) Puede haber estratos adicionales, lo cual parece ser suficiente para un trabajo óptimo respecto al costo de preparación.

En caso de duda respecto al uso del muestreo estratificado; o de formar o nó uno o varios estratos adicionales, la pauta que se sugiere es introducir el estrato adicional(a parte de lo que ya se tenga), sí y solo sí se está seguro que se obtendrá mayor ganancia en los resultados.

3) Si en la distribución la mayor parte de las características a ser medidas se restringe a pocas unidades muestrales, se aconseja "cortar" la cola de la distribución para formar un estrato y muestrearlo; puede haber casos en los que es preferible "partir" la distribución en partes y así formar 3 o más estratos. Algunas veces no es posible aplicar este principio debido primordialmente a restricciones administrativas. No obstante, se considera recomendable hacerlo(Román).

4) Generalmente no vale la pena formar un estrato pequeño a menos que sea en extremo heterogéneo con respecto a los otros estratos. En este caso,

los estimadores considerados son: media y desviación estándar(Deming).

La mejor caractrística para estratificar es la distribución de frecuencia de la variable - bajo estudio.

III.3.1 CONSTRUCCION DE ESTRATOS

Considérese una población de varianza finita distribuida en el intervalo (a,b) con una función- de densidad f, se estratifica la población de L estra - tos en y_h puntos de estratificación donde $a=y_o< y_1..<y_L=b$ por lo tento el h-ésimo estrato se forma con la subpo -- blación y_{h-1} y y_h.

Sean y_o, y_L el valor menor y mayor de y - en la población; se pretende encontrar límites interme - dios entre estratos $y_1, y_2, y_3 \ldots\ldots y_{L-1}$ tal que:

$$V(\bar{y}_c) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right) - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

sea mímima. Es suficiente minimizar $\sum W_h S_h$ si se ignora el coeficiente de variación.

Como $\bar{y}_h$ solo aparece en los términos - $W_h S_h$ y $W_{h+1} S_{h+1}$ al diferenciar tenemos:

$$\frac{\partial}{\partial y_h}\left(\sum W_h S_h\right) = \frac{\partial}{\partial y_h}\left(W_h S_h\right) + \frac{\partial}{\partial y_h}\left(W_{h+1} S_{h+1}\right)$$

Por otro lado tenemos que:

$$W_h = \int_{y_{h-1}}^{y_h} f(t)\,dt \quad , \quad \frac{\partial W_h}{\partial y_h} = f(y_h)$$

(donde y_h es la función de frecuencia de y).

$$W_h S_h^2 = \int_{y_{h-1}}^{y_h} t^2 f(t)\,dt - \frac{\left[\int_{y_{h-1}}^{y_h} t\, f(t)\,dt\right]^2}{\int_{y_{h-1}}^{y_h} f(t)\,dt}$$

diferenciando con respecto a y_h:

$$S_h^2 \frac{\partial W_h}{\partial y_h} + 2 W_h S_h \frac{\partial S_h}{\partial y_h} = y_h^2 f(y_h) - 2 y_h \mu_h f(y_h) + \mu_h^2 f(y_h)$$

(donde μ_h es la media de y en el estrato h).

Sumando $S_h^2 \frac{\partial W_h}{\partial y_h}$ al lado izquierdo y $S_h^2 f(y_h)$ al lado derecho de la igualdad, y dividiendo por $2 S_h$

$$\frac{\partial (W_h S_h)}{\partial y_h} = S_h \frac{\partial W_h}{\partial y_h} + W_h \frac{\partial S_h}{\partial y_h} = \frac{1}{2} f(y_h) \frac{(y_h - \mu_h)^2 + S_h^2}{S_h}$$

De la misma manera se obtiene:

$$\frac{\partial(W_{h+1}S_{h+1})}{\partial y_h} = -\frac{1}{2} f(y_h) \frac{(y_h-\mu_{h+1})^2+S^2_{h+1}}{S_{h+1}}$$

por lo tanto las ecuaciones para y_h son:

$$\frac{(y_h-\mu_h)^2+S^2_h}{S_h} = \frac{(y_h-\mu_{h+1})^2+S^2_{h+1}}{S_{h+1}}$$

(h = 1,2.....L-1)

Cuando S_h es la misma en todos los estratos, el total muestral se distribuye entre los estratos por asignación proporcional, bajo la cual el óptimp conjunto de estratos se obtiene al resolver - (L-1) ecuaciones simultáneas.

$$y_h = \frac{\mu_h+\mu_{h+1}}{2}$$

(h = 1,2.....L).

Dalenius y Gurney(1959) concluyeron - que dada $f(y)$, la regla es formar la cumulativa de $\sqrt{f(y)}$ y llegar al óptimo por aproximaciones.

$$\text{Sea} \quad Z(y) = \int_{y_0}^{y} \sqrt{f(t)}\, dt$$

$f(y)$ aproximadamente constante dentro de un estrato - dado si los estratos son numerosos y estrechos:

$$W_h = \int_{y_{h-1}}^{y_h} f(t)\, dt \doteq f_h (y_h - y_{h-1})$$

$$S_h = \frac{1}{\sqrt{12}} (y_h - y_{h-1})$$

$$Z_h - Z_{h-1} = \int_{y_{h-1}}^{y_h} \sqrt{f(t)}\, dt \doteq \sqrt{f_h}\, (y_h - y_{h-1})$$

donde f_h es el valor "constante" de $f(y)$ en el es - trato h . Sustituyendo obtenemos:

$$\sqrt{12} \sum W_h S_h \doteq \sum f_h (y_h - y_{h-1})^2 \doteq \sum (z_h - z_{h-1})^2$$

la suma en la derecha es minimizada haciendo - $(z_h - z_{h-1})$ constante.

Si el cumulativo sobre el rango total - (a,b) es H, las primeras aproximaciones a el óptimo - puntos de estratificación está definido por:

$$y_h = \frac{hH}{L} \qquad (h=1,2,.....L-1)$$

Dalenius, Hodges y Ekman (1959) concluyeron que los puntos y_h que satisfagan

$$W_h (y_h - y_{h-1}) \text{ constante}$$

proveerán aproximadamente los óptimos puntos de estratificación.

IV. CRITERIOS DE ASIGNACION DEL TAMAÑO DE MUESTRA EN ENCUESTAS PURAMENTE DESCRIPTIVAS

El principio a seguir en la determinación del tamaño de muestra es i) hacer uso efectivo de los recursos disponibles, ii) escojer la muestra de tal manera que proporcione un estimador de la media poblacional con precisión deseada a costo mínimo o iii) que proporcione un estimador con máxima precisión para un costo dado(Sukhatme,1970).

Cuando la asignación de la muestra en los diferentes estratos se realiza en base a dichos principios, obtenemos una Asignación Optima. La manera en la que el costo varíe con el tamaño de muestra total depende de la investigación particular. Como se sabe, el costo por experimento u observación, puede variar en los diferentes estratos dependiendo de la disponibilidad de los recursos y de la situación particular que se enfrente.

La elección de un tamaño de muestra consiste en términos generales en los siguientes pasos:

a) Debe haber una indicación relativa - a los límites de error deseado.

b) Encontrar una ecuación que conecte a n con la precisión deseada; dicha ecuación variará de acuerdo al diseño de muestreo planteado así como a la precisión.

c) La ecuación debe contener como parámetros ciertas propiedades de la población, las cuales son estimadas con el fin de dar resultados específicos.

d) Cuando se observa más de una característica y para cada una se especifica un grado de - precisión, los cálculos llevan a una serie de valores de n, uno para cada característica; por lo tanto, en este caso, se debe encontrar algún método para conciliar dichos valores, las operaciones se plantean i) asignando prioridades a las variables en estudio, ii)estableciendo una función multivariada y/o iii) determinando el tamaño de muestra como la n_i máxima del conjunto de n_i que satisfacen la precisión deseada para cada una de las caractrísticas en cuestión.

e) El paso final es considerar el valor de n y probar si es consistente con los recursos disponibles para elegir la muestra, es decir, estimación con referencia a costos de trabajo, tiempo, etc.

IV.1 MUESTREO ESTRATIFICADO PROPORCIONAL

En este tipo de estudio la fracción de muestreo es igual en todos los estratos; es decir, en cada estrato se obtiene una proporción igual de unidades de muestreo.

El muestreo estratificado proporcional es lo más cercano al concepto de "muestra representa - tiva", podemos omitir el término "representatividad" - debido a su carácter subjetivo; la representatividad - es un concepto ambiguo, una muestra es representativa se dice,"si representa a la población". Cualquier - muestra aleatoria o no aleatoria de una población por -ese mismo hecho la representa-. Si lo que se quiere - decir es que contiene todas las características rele - vantes de la misma, la posible conclusión es ambigua. De aquí que la "representatividad" tenga que asociarse a un concepto probabilístico, véase que el problema se ría "esta muestra es muy probablemente representativa" ó "tiene representatividad con probabilidad 0.75". En este sentido el enunciado "dice":

i) "Cubre a la población en el 75% de

los casos."

ii) "Están incluidos los elementos que caracterizan mayormente a la población en el 75% de los casos."

En la primera opción(i) se refiere al tamaño de muestra y así mientras mayor sea, mayor "representatividad" tendrá. En (ii) es subjetivo, ya que "¿Como sabemos sin utilizar fórmulas probabilísticas - que los elementos obtenidos "caracterizan mayormente" al 75% de la población?".

En suma, se recomienda omitir el término representativo por ser de carácter subjetivo. De aquí que varíe de investigación a investigación, de - muestra a muestra, y de usuario a usuario.

Sea

$$P_h = \frac{A_h}{N_h} \quad , \quad p_h = \frac{a_h}{n_h}$$

la proporción de unidades en X, en el h-ésimo estrato y en la muestra del estrato respectivamente. Para la proporción en la población total, el estimador es:

$$p_a = \frac{\sum N_h p_h}{N}$$

TEOREMA 1 Con el muestreo estratificado aleatorio la varianza de p_e es :

$$V(p_e) \doteq \frac{1}{N^2} \sum \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h}$$

En todas las aplicaciones aún si el coeficiente de variación no es despreciables, los términos $1/N_h$ serán despreciables y se puede usar la siguiente fórmula:

COROLARIO Cuando se puede ignorar el coeficiente de variación, se tiene:

$$V(p_e) \doteq \sum W_h^2 \frac{P_h Q_h}{n_h}$$

COROLARIO Con asignación proporcional:

$$V(p_e) = \frac{N - n}{N} \frac{1}{nV} \sum \frac{N_h^2 P_h Q_h}{N_h - 1}$$

$$\doteq \frac{1 - f}{n} \sum W_h P_h Q_h$$

IV.2 AFIJACION DE NEYMAN

El costo total se puede representar - por:

$$C = \sum c_i n_i \qquad (1)$$

donde c_i es el costo por observación o experimento en el i-ésimo estrato. Cuando c_i es el mismo de estrato - a estrato, digamos c, el costo total en una investiga ción está dada por:

$$C - cn$$

Consideremos la función:

$$\phi = V(\bar{y}_\omega) + \mu C$$

donde $\bar{y}_\omega$ es la media de la población y μ es al guna constante (Sukhatme).

$$V(\bar{y}_\omega) + \mu C = \sum\left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2 + \mu \left\{\sum c_i n_i\right\}$$

$$= \sum \left\{ \frac{p_i^2 S_i^2}{n_i} + \mu c_i n_i - 2 p_i S_i \sqrt{\mu c_i} + \right.$$

$$\left. + 2 p_i S_i \sqrt{\mu c_i} - \frac{1}{N_i} p_i^2 S_i^2 \right\}$$

$$= \sum \left(\frac{p_i S_i}{\sqrt{n_i}} - \sqrt{\mu c_i n_i} \right)^2 + 2 \sum p_i S_i \sqrt{\mu c_i}$$

$$- \sum \frac{1}{N_i} p_i^2 S_i^2$$

$$= \sum \left(\frac{p_i S_i}{\sqrt{n_i}} - \sqrt{\mu c_i n_i} \right)^2 \qquad (2)$$

$V(\bar{y}_\omega)$ es mínima para C fijo, o el costo de una investigación es mínimo para un valor fijo de $V(\bar{y}_\omega)$ cuando cada uno de los términos cuadráticos del segundo término de (2) es cero, o en otras palabras, cuando

$$n_i \propto \frac{p_i S_i}{\sqrt{c_i}} \qquad (i = 1, 2 \ldots k) \qquad (3)$$

Esta ecuación nos muestra:

i) Mientras mayor sea el tamaño del estrato, mayor debe ser la contribución del estrato en la muestra que se va a seleccionar(estrato proporcional).

ii) Entre mayor sea la variabilidad dentro del estrato, mayor debe ser el tamaño de la muestra para este estrato.

iii) Entre más barato sea el costo en un estrato, mayor debe ser la muestra en ese estrato.

Para satisfacer la condición de costo o varianza fija, evaluamos la constante de proporcionalidad $\frac{1}{\sqrt{u}}$.

Sustituyendo (3) en (1) obtenemos:

$$C_o = \sum \frac{p_i S_i}{\sqrt{u c_i}} c_i$$

C_o cantidad presupuestada, con la cual se desea estimar la media con la máxima precisión, por lo tanto:

$$\frac{1}{\sqrt{u}} = \frac{C_o}{\sum p_i S_i \sqrt{c_i}}$$

de donde:

$$n_i = \frac{p_i S_i C_o}{\sqrt{c_i}\left(\sum p_i S_i \sqrt{c_i}\right)}$$

Cuando $c_i = c$ (i=1,2....k). El costo - de investigación es proporcional al tamaño de muestra, las n_i están dadas por:

$$n_i = n \frac{p_i S_i}{\sum p_i S_i}$$

La afijación con esta fórmula nos pro_ porciona la estimada de la media con precisión máxima para un tamaño dado de muestra.

Cuando la media de la población se desea estimar con una varianza dad, digamos V_o, con cos_ to mínimo, evaluamos la constante de proporcionali - dad sustituyendo a n_i de (3) en:

$$\sum \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_1^2 = V_o$$

y obtenemos, puesto que $p_i = N_i / N$,

$$\frac{1}{\sqrt{n}} = \frac{\sum p_i S_i \sqrt{c_i}}{V_o + \frac{1}{N}\sum p_i S_i^2}$$

De donde:

$$n_i = \frac{p_i S_i}{\sqrt{c_i}} \cdot \frac{\sum p_i S_i \sqrt{c_i}}{V_o + \frac{1}{N} \sum p_i S_i^2}$$

Cuando $c_i = c$ la ecuación anterior se reduce a :

$$n_i = p_i S_i \frac{\sum p_i S_i}{V_o + \frac{1}{N} \sum p_i S_i^2}$$

de manera que la muestra mínima requerida para esti - mar la media con varianza fija V_o está dada por:

$$n = \frac{\left(\sum p_i S_i\right)^2}{V_o + \frac{1}{N} \sum p_i S_i^2}$$

IV.3 ESTIMACION DEL TAMAÑO DE MUESTRA CON INFORMACION CONTINUA

En esta sección se presentan fórmulas para cualquier asignación, si se supone que el estimador tiene una varianza V especificada, si en su lugar el margen de error d ha sido especificado, entonces la varianza es $V=\left(\frac{d}{t}\right)^2$ en donde t es el desvío normal correspondiente a la probabilidad permisible de - que el error excederá el margen deseado(Cochran,1976).

ESTIMACION DE LA MEDIA DE LA POBLACION Y

Sea s_h el estimador de S_h y sea $n_h=w_h n$ donde las w_h han sido seleccionadas, entonces:

$$V = \frac{1}{n}\sum \frac{W_h^2 S_h^2}{w_h} - \frac{1}{N}\sum W_h S_h^2$$

donde $W_h = N_h / N$ por lo tanto para n tenemos que:

$$n = \frac{\sum \frac{W_h \Delta_h^2}{w_h}}{V + \frac{1}{N}\sum W_h \Delta_h^2}$$

si se ignora el coeficiente de variación, se tiene como primera aproximación:

$$n_o = \frac{1}{V}\sum \frac{W_h^2 \Delta_h^2}{w_h}$$

por otro lado, si n_o / N no es despreciable, entonces n la obtenemos de la siguiente manera:

$$n = \frac{n_o}{1 + \frac{1}{NV}\sum W_h \Delta_h^2}$$

Considerando asignación óptima(para n - fija):

$$w_h \propto W_h s_h$$

$$n = \frac{(\sum W_h \Delta_h)^2}{V + \frac{1}{N}\sum W_h \Delta_h^2}$$

Con asignación proporcional:

$$w_h = W_h = N_h / N$$

$$n_o = \frac{\sum W_h \Delta_h^2}{V} \qquad n = \frac{n_o}{1 + \frac{n_o}{N}}$$

ESTIMACION TOTAL DE LA POBLACION

Si V es la deseada $V(\hat{Y}_e)$, las principales fórmulas son:

$$n = \frac{\sum N_h^2 s_h^2 / \omega_h}{V + \sum N_h s_h^2}$$ General

$$n = \frac{(\sum N_h s_h)^2}{V + \sum N_h s_h^2}$$ Supuesto Optimo (n fija)

$$n_o = \frac{N}{V} \sum N_h s_h^2 \quad , \quad n = \frac{n_o}{1 + \frac{n_o}{N}}$$ Proporcionaal

IV.4 ESTIMACION DEL TAMAÑO DE MUESTRA CON PROPORCIONES

Sea V la varianza deseada en el estimador de la proporción P en la población total. Las fórmulas para los dos tipos principales de asignación son los siguientes:

$$n_0 = \frac{\sum W_h p_h q_h}{V} \quad , \quad n = \frac{n_0}{1 + \frac{n_0}{N}} \qquad \text{Proporcional}$$

$$n_0 = \frac{\left(\sum W_h \sqrt{p_h q_h}\right)^2}{V} \quad , \quad n = \frac{n_0}{1 + \frac{1}{NV} \sum W_h p_h q_h} \qquad \text{Optimo supuesto}$$

donde n_0 es la primera aproximación, la cual ignora la fracción de muestreo , y n es el valor corregido considerando fracciones de muestreo . En el desarrollo de las fórmulas, los factores N_h / N_{h-1} son considerados como la unidad(Cochran).

V. CRITERIOS DE ASIGNACION DEL TAMAÑO DE MUESTRA EN ENCUESTAS ANALITICAS

El propósito principal al trabajar con encuestas analíticas, es hacer comparaciones entre diferentes estratos. Las reglas para asignar los tamaños de muestra son diferentes de las que se usan cuando el objetivo es obtener estimadores de la población como un todo.

Las subpoblaciones o dominios de estudio denotan subclases, tales como regiones, controladas de manera específica en el tamaño muestral. La mayor parte de las subclases no se puede asignar fácilmente por lo que se aceptan tal como se encuentran, - por ejemplo, las clases de edad y sexo en muestras de vivienda; por otro lado, para cada dominio se planea la encuesta para proporcionar mayor información.

Se consideran dominios de estudio, áreas locales, unidades administrativas, áreas geográficas.

La clasificación de dominios de manera general es la siguiente:

1) Dominios "planeados. Son dominios - para los cuales se ha planeado, designado y seleccionado estimaciones por separado en el diseño de muestra, su combinación forma la muestra total, por ejemplo, regiones mayores y estratos individuales compuestos de unidades primarias.

2) "Clases Cruzadas o Clases Transversales". Es una combinación entre diseño muestral y las unidades muestrales, por ejemplo, edad, sexo, ocupación, y niveles de educación.

3) Entre los dos anteriores se encuentran las "divisiones", son poco usadas, no se diferencian en la selección muestral, pero tienden a concentrarse de manera desigual en unidades primarias.

El tamaño de los dominios influye en la selección de métodos a seguir, por otro lado, en algunos dominios es necesario que la fracción de muestreo aumente para obtener la precisión deseada.

Los tamaños de dominios a manera general es:

a) Dominios Mayores. Comprenden 1/10 - ó más de la población, por ejemplo, regiones mayores, grupos por edad(10 años), o clases de categorías mayores, tales como ocupaciones.

b) Dominios Menores. Comprenden entre - 1/10 y 1/100 de la población, por ejemplo, clasificaciones dobles tales como ocupación por educación; clasificación unitaria, por ejemplo, incapacidad de trabajo.

c) Mini Dominios. Constan entre 1/100 y 1/10,000 de la población, como ejemplos tenemos, poblaciones de provincias, clasificaciones triples, tales como: edad con ocupación con educación.

d) Tipos poco comunes de individuos. Constan de menos de 1/10,000 de la población, como ejemplos tenemos: poblaciones de áreas de servicio médico clasificadas según grupos étnicos; e individuos con problemas de salud crónica, clasificados por áreas de residencia.

Cabe señalar que los tamaños aquí mencionados son aproximados; el tamaño realmente depende del tamaño de las muestras y poblaciones, de las va -

riables y estadísticas, de la precisión y decisión, - etc.

Los métodos más usados son los siguientes:

i) Técnicas de Conteo Sintomático (SAT) son las técnicas más antiguas, relacionan el crecimiento en la población al crecimiento de variables "sintomáticas", tales como números de nacimientos y muertes, habitación, etc. usan principalmente relaciones demográficas combinadas con relaciones estadísticas.

ii) Estimación Sintética, usa datos - muestrales para estimar en algún nivel la variable de interés para diferentes subclases de la población, entonces mide estos estimadores en proporción a la incidencia de subclases dentro de los dominios(pequeños) - de interés, por ejemplo, estimaciones de desempleo, - tablas cruzadas por edad, sexo y raza, pueden ser medidas por la incidencia proporcional de estas clases en cada provincia para estimar el desempleo por pro-- vincias.

iii) Procedimientos de Regresión Sintomática, consiste en ajustar una relación funcional en-- tre las variables de interés y las variables sintomá-

ticas, se usan principalmente para estimación de las - variables continuas tales como ingresos económicos.

iv) Métodos de Regresión Muestral, están basados en ecuaciones de regresión usando variables - sintomáticas(como indicadores), medidas para cada do— minio como variables independientes, por otro lado los datos muestrales para las variables de interés se consideran variables dependientes.

v) Métodos de base unitaria, este método consiste en dividir los dominios pequeños en unidades de base más pequeñas, después clasificar cada una de estas"bases unitarias" en una de k grupos de bases unitarias para las cuales los estimadores se pueden - obtener de los conocidos métodos muestrales.

Se consideran unidades de base: bloques, enumeración de distritos, provincias u otras áreas - geográficas(Kish y Purcell).

NOTACION (Para unidades en el estrato h, contenidas en el dominio j)(Cochran)

Número de unidades	N_{hj} $\quad \sum_j N_{hj} = N_h$
Número en la muestra	n_{hj} $\quad \sum_j n_{hj} = n_h$
Medición en la unidad individual	y_{hcj}
Media de la muestra	$\bar{y}_{hj} = \sum_{i=1}^{n_{hj}} \frac{y_{hcj}}{n_{hj}}$
Media del dominio	$\bar{Y}_{hj} = \sum_{i=1}^{N_{hj}} \frac{y_{hcj}}{N_{hj}}$
Total poblacional	$Y_j = \sum_h N_{hj} \bar{Y}_{hj}$
Media para el dominio j sobre todos los estratos	$\bar{Y}_j = \frac{Y_j}{N_j}$, $N_j = \sum_h N_{hj}$
Estimador de Y_j	$\hat{Y}'_j = \sum_h N_{hj} \bar{y}_{hj}$

Estimador de $\bar{Y}_j$

$$\hat{\bar{Y}}_j = \frac{\hat{Y}_j}{N_j}$$

Varianza de $\hat{Y}_j$.

$$V(\hat{Y}_j) = \sum_h \frac{N_{hj}^2 S_{hj}^2}{n_{hj}} \left(1 - \frac{n_{hj}}{N_{hj}}\right)$$

donde S_{hj}^2 es la varianza entre unidades en el dominio j dentro del estrato h; nótese que las N_{hj} no siempre se conocen en la práctica.

V.I ESTIMACION DE LOS TOTALES DE LOS DOMINIOS

Para obtener un estimador del total - del dominio, se calcula el estimador total por subpoblación, se suman, y de esta manera tenemos la estimación buscada, es decir: Si no son conocidas la N_{hj} - multiplicamos el total muestral y_{hj} de unidades que - están en el j-ésimo dominio, y el estrato h por el factor de expansión N_h / n_h , esto dá el estimador(por dominio):

$$\hat{Y}_j = \frac{Nh}{n_h} \sum y_{hj}$$

ahora sumando los dominios tenemos el estimador:

$$\hat{Y}_j = \sum \frac{Nh}{n_h} \sum y_{hij}$$

La varianza estimada es:

$$v(\hat{Y}_j) = \sum \frac{N_h^2}{n_h(n_h-1)} (1-f_h) \left[\sum y_{hij}^2 - \frac{(\sum y_{hij})^2}{n_h} \right]$$

V.2 ESTIMACION DE LAS MEDIAS DE DOMINIOS

Para estimar la media del dominio Y_j/N_j se requiere un estimador de N_j de la muestra. Un estimador insesgado es:

$$\hat{N}_j = \sum \frac{N_h}{n_h} n_{hj}$$

Tomemos:

$$\hat{\bar{Y}}_j = \frac{\hat{Y}_j}{\hat{N}_j} = \frac{\sum_h \frac{N_h}{n_h} \sum_i y_{hij}}{\sum_h \frac{N_h}{n_h} n_{hj}}$$

Con asignación proporcional $\hat{\bar{Y}}_j$ se reduce a la media ordinaria de la muestra de las unidades que caen en el dominio j

Sea x'_{hi} (variable muda), la cual es igual a 1 para todas las unidades en el dominio j y 0 para todas las demás unidades(donde i = 1.....N_h), por lo tanto:

$$\bar{x}_h = \frac{\sum_i^{n_h} x'_{hi}}{n_h} = \frac{n_{hj}}{n_h}$$

$$\bar{y}'_h = \frac{\sum^{n_h} y_{hi}}{n_h} = \frac{\sum^{n_{hj}} y_{hij}}{n_h} = \frac{n_{hj}}{n_h} \bar{y}_{hj}$$

y la media estimada del dominio se puede escribir:

$$\hat{\bar{Y}}_j = \frac{\sum_h \frac{N_h}{n_h} \sum y_{hij}}{\sum \frac{N_h}{n_h} n_{hj}} = \frac{\sum_h N_h \bar{y}'_h}{\sum N_h x'_h} = \frac{\bar{y}'_e}{\bar{x}'_e}$$

la cual es un estimador de razón combinado para las - variables y'_{hi} y x'_{hi}.

La varianza estimada se puede expresar aproximadamente como:

$$v(\hat{\bar{Y}}_j) = \frac{1}{\hat{N}_j^2} \sum_h \frac{N_h^2 (1-f_h)}{n_h (n_h-1)} \sum_i^{n_h} \left[y'_{hi} - \hat{\bar{Y}}_j x'_{hi} - (\bar{y}'_h - \hat{\bar{Y}}_j \bar{x}'_h) \right]^2 \qquad (1)$$

La segunda suma se puede escribir como:

$$\sum_i^{n_h} (y'_{hi} - \hat{\bar{Y}}_j x'_{hi})^2 - n_h (\bar{y}'_h - \hat{\bar{Y}}_j \bar{x}'_h)^2 = \sum_i^{n_{hj}} (y_{hij} - \hat{\bar{Y}}_j)^2 - \frac{n_{hj}^2}{n_h} (\bar{y}_{hj} - \hat{\bar{Y}}_j)^2$$

El primer término se puede expresar:

$$\sum_i^{n_{hj}} (y_{hij} - \bar{y}_{hj})^2 + n_{hj} (\bar{y}_{hj} - \hat{\bar{Y}}_j)^2$$

sustituyendo en (1) tenemos:

$$v(\hat{\bar{Y}}_j) = \frac{1}{\hat{N}_j^2} \sum_h \frac{N_h^2 (1-f_h)}{n_h (n_h-1)} \left[\sum_i (y_{hij} - \bar{y}_{hj})^2 + n_{hj} \left(1 - \frac{n_{hj}}{n_h}\right) (\bar{y}_{hj} - \hat{\bar{Y}}_j)^2 \right]$$

que es la varianza estimada(Cochran,1976).

V.3 METODO DE ESTIMACION SINTETICA

Supóngase que se desea estimar una característica x dentro de dominios pequeños, y también tenemos información sobre algunas variables asociadas Y, las cuales son no "traslapadas" y son subgrupos exhaustivos de la población.

Es decir tenemos la siguiente información, el número de variables asociadas al h-ésimo dominio y el g-ésimo grupo (Y_{hg}) y la estimación de la característica x para el subgrupo g, a nivel de "dominio mayor" basada en la muestra (x'_{ig}), por lo tanto, el estimador sintético total de la característica x en el dominio pequeño h es:

$$\hat{X}_h = \sum_g \hat{X}_{hg} = \sum_g \left(\frac{Y_{hg}}{Y_{ig}} \right) x'_{ig} \qquad (1)$$

Sea $Y_{hg} = N_{hg}$ (ya que la variable asociada Y, por lo tanto, (1) se convierte en :

$$\hat{X}_n = \sum_g \left(\frac{N_{ng}}{N_{ig}}\right) x'_{ig}$$

Es deseable que éste último corresponda a un estimador de dominio mayor cuando al sumarse sobre dominios exhaustivos pequeños tenemos:

$$\sum_n \hat{X}_n = \sum_n \sum_g \left(\frac{Y_{ng}}{Y_{ig}}\right) x'_{ig}$$

$$= \sum_n \sum_g \left(\frac{Y_{ng}}{Y_{ig}}\right) x'_{ig}$$

$$= \sum_g x'_{ig} = X''$$

Sea

$$\bar{X}_n = \sum_g \left(\frac{Y_{ng}}{Y_{nj}}\right) \bar{X}'_{ig} \qquad (2)$$

el estimador sintético de la media del dominio pequeño para la característica x y $\bar{x}_{ig}$ el estimador de la investigación de la media de la característica x del subgrupo g a nivel de dominio mayor.

Para ambos estimadores se usa Y_{hg}, tamaño de la variable asociada para dominios pequeños , sin embargo, (2) se usa en los subgrupos dentro de dominíos pequeños, mientras que en (1) es usado en dominios pequeños dentro de los subgrupos.

La estimación sintética reduce varianzas, son estimadores insesgados por dos razones:

i) Con frecuencia existen "desviaciones" de los supuestos fundamentales de homogeneidad de proporciones y

ii) Las ponderaciones Y_{hg} / Y_{ig} se basan generalmente en datos previos y la estructura de la población puede cambiar, por lo tanto la expresión del sesgo para (1) es:

$$E\left[\hat{x}_h - X_h\right] = E\left[\sum_g \left(\frac{Y_{hg}}{Y_{ig}}\right) x'_{ig} - \sum_g X_{hg}\right]$$

$$= \sum_g Y_{hg}\left(\frac{X_{ig}}{Y_{ig}} - \frac{X_{hg}}{Y_{hg}}\right)$$

por lo tanto el estimador es segado para X_h a menos - que $X_{ig} / Y_{ig} = X_{hg} / Y_{hg}$ para todos los grupos ($Kish_1$).

Sin embargo, esto no es cierto siem — pre. Es complicado evaluar los estimadores sintéticos debido al sesgo, por lo tanto, ponemos especial atención en el error cuadrático medio, aún esta estima — ción es difícil debido a la falta de información de - X_{hg} y son valores que se necesitan para estimar el sesgo; González y Waskberg (1973) sugieren el uso de el promedio sobre los dominios de interés; el cual está dado por:

$$\sum_h \frac{E\left(\hat{x}_h - X_h\right)^2}{H}$$

donde H es el número de dominios pequeños, además el - estimador de este promedio se deriva de suposiciones - limitadas.

Una de las dificultades al trabajar - con estos estimadores, es que a menos que las varia - bles x-s están altamente correlacionadas con las variables de interés, los estimadores sintéticos tienden a agruparse alrededor de la media para el dominio mayor, lo cual se refleja principalmente en factores de áreas locales. Para estimadores pequeños la selección cuidadosa de las variables proporciona resultados útiles. (Kish y Purcell).

V.4 ESTRATOS COMO DOMINIOS DEL ESTUDIO

Pondremos el caso en que sólo tenemos dos estratos, podemos escojer n_1 y n_2 para minimizar la varianza de la diferencia $(\bar{y}_1 - \bar{y}_2)$ entre las medias estimadas de los estratos, es decir:

$$V(\bar{y}_1 - \bar{y}_2) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

para la cual se tiene la función de costo lineal:

$$C = c_0 + c_1 n_1 + c_2 n_2$$

por lo tanto afirmamos que V es minimizada cuando:

$$n_1 = \frac{\frac{nS_1}{\sqrt{c_1}}}{\frac{S_1}{\sqrt{c_1}} + \frac{S_2}{\sqrt{c_2}}} \qquad n_2 = \frac{\frac{nS_2}{\sqrt{c_2}}}{\frac{S_1}{\sqrt{c_1}} + \frac{S_2}{\sqrt{c_2}}} \qquad (1)$$

Con L estratos, L 2, la asignación - óptima depende de la precisión deseada para las diferentes comparaciones. Por ejemplo, el costo puede ser minimizado sujeto al grupo de L(L-1)/2 condiciones que $V(\bar{y}_h - \bar{y}_i)$ V_{hi}, donde los valores de V_{hi} se esco-

jen de acuerdo con la precisión considerada necesaria para una comparación satisfactoria de los estratos h e i.

Si las S_h y c_h no difieren grandemente; un método es minimizar la varianza promedio de la diferencia entre todos los L(L-1)/2 pares de estratos, esto es minimizar:

$$V = \frac{2}{L}\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \cdots\cdots + \frac{S_L^2}{n_L}\right)$$

V es minimizada, para C fijo mediante (1).

$$n_h \propto S_h / c_h$$

VI. CRITERIOS DE ASIGNACION EN ENCUESTAS ANALITICAS (CASO HIPOTESIS ESPECIFICAS SOBRE LOS PARAMETROS DE LA POBLACION.)

Por su sencillez, los métodos de muestreo se han desarrollado alrededor de la teoría univariada. Sin embargo, hay muchas encuestas(y sus muestras correspondientes), quizá casi todas, que tienen multiplicidad de objetivos; se puede observar que las encuestas con objetivos múltiples tienen ventajas tales como:

1) Se pueden medir varias características(variables de encuesta) con el mismo conjunto de elementos muestrales; dichas ventajas son similares a las del diseño experimental de factores múltiples; es decir:

i) Hay un menor costo;

ii) Mayor alcance; pues además de evaluar los efectos de los factores aislados, se evalúan sus interacciones posibles.

2) Las observaciones repetidas en los mismos elementos muestrales, se pueden considerar como casos singulares de variables múltiples. La correlación, así entre distintas variables puede obtenerse con oportunidad y mayor precisión.

3) Sirve para determinar las diferentes subclases por medio de las variables adecuadas a la encuesta; la relación entre las características de la encuesta es análoga a la relación de las variables dependientes de las independientes en modelos lineales, por ejemplo, pueden darse estadísticas por regiones, edades, sexo, ocupaciones, niveles de educación y muchas otras clases, así como sus combinaciones.

En las encuestas múltiples se tienen dos tipos de suposiciones respecto a la disponibilidad de datos(en este enfoque se puede utilizar una amplia variedad y formas de información de las variables que están relacionadas con las variables de interés, por ejemplo, edad, sexo, raza, que están ampliamente vinculados con empleos en áreas locales. Estas "variables asociadas" quienes dividen con efectividad la población en "clases cruzadas o transversales" tienen gran importancia).

i) Se necesita una "estructura asociada", es decir, establecer en algún punto previo(que puede see el valor de lagunas variables obtenidas en otra encuesta o provista por otra fuente(empírica) o vía un marco de referencia teórico) la relación entre la(s) variable(s) de interés y las "variables asociadas"(esto se hace a nivel de dominios pequeños).

ii) Necesitamos una "estructura de asignación", es decir, establecer relaciones entre la variable de interés y las variables asociadas(a nivel dominio mayor) acumuladas en los dominios pequeños bajo estudio.

El procedimiento es entonces ajustar la estructura asociada con los datos de la estructura de asignación y/o al mismo tiempo mantener las relaciones entre las variables de la estructura asociada.

El punto importante de este enfoque es que el proceso de estimación está completamente especificado por las dos estructuras de datos(asociada y de asignación). En conclusión, en este procedimiento tenemos que:

a) Ajustar la información disponible en la estrustura de asignación y

b) Conservar de alguna manera las relaciones presentes en la estructura de asociación sin - interferir con (a).

Por otro lado, la forma de los estimadores resultantes que satisfagan (a) y (b) dependerán de la información involucrada en las estructuras mencionadas.

En el contexto estadístico clásico, una hipótesis, es un supuesto o conjetura concerniente a - una población bajo estudio. Las pruebas de hipótesis - se establecen comúnmente en base a las distribuciones de estadística útiles y que se desprenden del modelo (Z, t, χ^2, y F).

Los pasos fundamentales empleados en el procedimiento según los diferentes métodos para probar hipótesis son:

1) Establecer las hipótesis "nulas", - por ejemplo, en las siguientes formas: i) No hay diferencia entre dos valores dados , o ii) la diferencia es cero. En otras palabras, se hace el supuesto que - la diferencia entre los dos valores dados es solamente debida a fluctuaciones en el muestreo o al azar; por

lo tanto, la diferencia es considerada como "no diferencia" o "no significativa".

2) Expresar la diferencia en unidades del error estándar del estimador de la siguiente manera:

i) Cuando la verdadera desviación estándar de la población σ es conocida, el error estándar del estimador es conocido, por lo tanto, la diferencia es expresada en el valor de la desviación normal estándar Z tal como:

$$Z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

ii) Cuando la verdadera desviación estándar de la población σ es desconocida y el error estándar del estimador, es calculado a partir de una muestra, t , la diferencia es espresada en el valor de t tal como:

$$t = \frac{\bar{x}-\mu}{s_{\bar{x}}} = \frac{\bar{x}-\mu}{s/\sqrt{n-1}}$$

cuando el valor de t es obtenido, se usan tablas de distribución en las pruebas de hipótesis.

3) Tomar una decisión. En el enfoque clásico, la regla de decisión está basada en el nivel de significación de la prueba. i) Si el valor calculado de la estadística de prueba cae en la región de aceptación, consideramos la diferencia entre los dos valores propuestos como por ejemplo debida a la fluctuación de muestreo o a contribución de variables no incluidas en la prueba de hipótesis, y por lo tanto no rechazamos la hipótesis. ii) Si el valor calculado de la estadística de prueba cae en la región de rechazo, consideramos la diferencia entre los valores propuestos como no debida a la fluctuación de muestreo y rechazamos la hipótesis.

En el caso de poblaciones finitas, como se sabe, no se presupone una distribución específica -si bien y de acuerdo a los teoremas límite de la probabilidad- se puede hacer uso de distribuciones aproximadas. Dicho punto, establece, entonces, diferencias.

Para ilustrar el problema de asignación en encuestas del tipo analítico se discuten los siguientes casos:

Considérese una población dividida en dos estratos E_1 y E_2 que actúan, a la vez, como domi-

nios de estudio. Aquí se requiere comparar parámetros o valores poblacionales de los estratos.

El caso es relativamente común, por ejemplo, i) se desean comparar dos regiones en su consumo o en su producción; ii) se tiene como propósito comparar el rendimiento y/o "la eficiencia" de dos modalidades educativas, iii) se busca evaluar si existe diferencia en productividad en dos distintas organizaciones industriales, etc.

Así una pregunta inicial consiste en la asignación de un tamaño de muestra n(previamente obtenido) en los dos estratos.

Al ser una Encuesta Analítica con una hipótesis de comparación la situación puede ser formalizada vía:

1) $H_o : \theta_1 = \theta_2$ vs. $H_A : \theta_1 = \theta_2$

2) $H_o : \theta_1 = \theta_2$ vs. $H_A ; \theta_1 > \theta_2$

3) $H_o : \theta_1 = \theta_2$ vs. $H_A : \theta_1 < \theta_2$

(según la información con que se cuente y el propósito de la investigación).

θ actúa aquí como representación de cualquier parámetro o, para utilizar el enfoque de Kish, como cualquier valor poblacional. De esa manera θ puede ser un un total: Y; un valor promedio: $\bar{Y}$, una proporción: P, una razón: Y/Z, o en la contribución de varias variables en un modelo de regresión $Y_i = \alpha + \beta x_i$.

Se revisa aquí, inicialmente, $\theta = \bar{Y}$ en una prueba de dos colas.
Entonces,

$$H_o : \bar{Y}_1 = \bar{Y}_2 \quad \text{vs.} \quad H_A : \bar{Y}_1 \neq \bar{Y}_2$$

Es decir, se establece como hipótesis nula que el promedio(de ingresos, consumo, escolaridad, horas de exposición a la televisión o de afiliados a algún partido político) de la característica en cada estrato es similar(o igual). La hipótesis alternativa considera los casos en donde existen diferencias significativas en cada estrato.

De aquí que n_1 y n_2 se obtengan en base a:

1) Se establece la estadística de prueba(estimador):

Como $\bar{y}_1 = \hat{\bar{Y}}_1$ y $\bar{y}_2 = \hat{\bar{Y}}_2 \Longrightarrow$

$z = \bar{y}_1 - \bar{y}_2$ contiene la información relevante para la conducción de la prueba. Como se visualiza las propiedades de z son:

$$E(z) = \bar{Y}_1 - \bar{Y}_2$$

$$H_0, E(z) = 0$$

$$H_A, E(z) \neq 0$$

$$Var(z) = Var(\bar{y}_1 - \bar{y}_2)$$

$$Var(z) = Var\,\bar{y}_1 + Var\,\bar{y}_2 - 2\,Cov(\bar{y}_1, \bar{y}_2)$$

$$Var(z) = Var\,\bar{y}_1 + Var\,\bar{y}_2 \;;$$

$\bar{y}_1, \bar{y}_2$ se obtienen por muestreo estratificado con muestreo simple aleatorio en cada estrato.

Es decir,

$$Var(z) = (1 - f_1)\frac{S_1^2}{n_1} + (1 - f_2)\frac{S_2^2}{n_2}$$

$$\left(f_i = \frac{n_i}{N_i}\right)$$

$$Var(z) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + K$$

$$\left[K = -\left(S_1^2/N_1 + S_2^2/N_2\right)\right.$$

K actúa como constante ya que no depende de las variables de decisión n_1 y n_2. Por lo tanto, la función a minimizar(para n_1 y n_2) dentro del problema de asignación es:

$$Var(\bar{y}) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

$$s.a. \quad n = n_1 + n_2$$

De aquí se obtiene un problema de Programación Matemática Entera. Una solución rápida se deriva de la siguiente manera:

Sea $$\tilde{S} = \min\{S_1, S_2\}$$

Por ejemplo $$\tilde{S} = S_1 \Rightarrow \exists\ k>0 \ \ni\ S_2 = kS_1$$

Consecuentemente, sin considerar la restricción $n_2 = kn_1$ (n_2 mayor que n_1) y n_1 se obtiene mediante:

$$n = n_1 + n_2$$

$$n = (1+k)n_1$$

$$\Rightarrow \quad n_1 = \frac{n}{(1+k)} \ ; \ si \ \frac{(1+k)}{n} \ y \ n_2 = kn_1 \in \mathbb{Z}^+$$

ya se obtuvo una solución factible(de hecho, la óptima). Si no es así, una solución aproximada por redoendeo pue

de servir. La mayoría de las veces aumentando el valor de n. En conclusión n_1 y n_2 no serán necesariamente iguales a menos que sus cuasivarianzas(S_1 y S_2) lo sean.

El siguiente caso refleja dos propósitos en la Encuesta por Muestreo:

(1) Estimar $\bar{Y}$ total para los dos estratos

es decir, estimar $$\bar{Y}_a = W_1\bar{Y}_1 + W_2\bar{Y}_2$$

(2) Comparar los dos estratos(dominios de estudio) en términos de $\bar{Y}_1$ vs. $\bar{Y}_2$

$$H_o : \bar{Y}_1 = \bar{Y}_2 \quad vs. \quad H_A : \bar{Y}_1 \neq \bar{Y}_2$$

Para (1) la asignación de n nos queda $(\bar{y}_a = \hat{\bar{Y}}_e)$ en base a:

$$\min_{n_1, n_2} \; Var(\bar{y}_e) = W_1^2 Var(\bar{y}_1) + W_2^2 Var(\bar{y}_2)$$

Para (2) la asignación de n es equivalente a la solución de:

$$\min_{n_1, n_2} \; \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

es decir, se establecen dos funciones para minimizar:

$$f = W_1^2 (1-f_1) \frac{S_1^2}{n_1} + W_2^2 (1-f_2) \frac{S_2^2}{n_2} \sim W_1^2 \frac{S_1^2}{n_1} + W_2^2 \frac{S_2^2}{n_2}$$

$$g = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

en f interviene explícitamente el tamaño de cada estrato por separado. La minimización de f se daría vía $n_i \propto W_i^2 S_i^2$ y la de g vía $n_i \propto S_i^2$. La conjunta(si ambos propósitos tienen la misma importancia) sería vía $n_i \propto (1+W_i^2) S_i^2$. En general, la minimización se establece vía:

$$h = \alpha \left(W_1^2 \frac{S_1^2}{n_1} + W_2^2 \frac{S_2^2}{n_2} \right) + \beta \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)$$

$$\alpha + \beta = 1$$

$$\alpha \geq 0 \; ; \; \beta \geq 0$$

Como se puede observar en los casos anteriores, la complejidad de la asignación aumenta conforme se inscriben más hipótesis o la hipótesis se refiere a un parámetro con estimadores más complejos, como el obtenido bajo diseños que incluyen varias etapas autoponderadas o no.

La situación mencionada puede hacerse más evidente si se consideraran en los dos estratos E_1 y E_2 estimadores de razón. Dicho caso corresponde a la comparabilidad de por ejemplo, i) tasas de cos--

tos diferenciales, ii) variables correlacionadas en - el consumo(sea R = Y/X donde Y: litros de leche consumidos en las unidades familiares y X: consumo correspondiente de litros de refresco, iii) en general, variables correlacionadas o de las que busque establecer niveles de correlación:

$$y' = \bar{y} + k(\bar{x} - \bar{x})$$

$$k = \frac{\bar{y}}{\bar{x}}$$

$k = b$ k constante

De aquí que la representación fuese:

$R_1 = Y_1/X_1$ (para el estrato E_1)

$R_2 = Y_2/X_2$ (para el estrato E_2)

entonces: $H_o : R_1 = R_2$ vs. $H_A : R_1 \neq R_2$

que es idéntica a:

$H_o : Y_1/X_1 = Y_2/X_2$ vs. $H_A : Y_1/X_1 \neq Y_2/X_2$

como $\hat{R}_1 = \frac{y_1}{x_1}$, $\hat{R}_2 = \frac{y_2}{x_2}$ $(\hat{R}_i \cdot 7 \cdot E(\hat{R}_i) \neq R_i)$

Entonces, $z = y_1/x_1 - y_2/x_2$ no es estimador insesgado de $Z = R_1 - R_2$.

De aquí que $E[z/H_0]$ sea tan sólo aproximadamente cero($E[\hat{R}_1] = E[r] \doteq R + \frac{1-f}{n\bar{x}_2}[RS_2^2 - \rho S_x S_y]$

Para minimizar $Var[z] = Var(y_1/x_1) + Var(y_2/x_2)$ se observa que:

$$Var[z] = Var(r_1) + Var(r_2)$$

Como $$Var(r_1) = \frac{1}{\bar{x}_1^2}\left[V(\bar{y}_1) + R_1^2 V(\bar{x}_1) - 2R_1 Cov(\bar{x}_1, \bar{y}_1)\right.$$

$$Var(r_2) = \frac{1}{\bar{x}_2^2}\left[V(\bar{y}_2) + R_2^2 V(\bar{x}_2) - 2R_2 Cov(\bar{x}_2, \bar{y}_2)\right.$$

$$\Rightarrow \quad Var(z) = \sum_{i=1}^{2} \frac{1}{\bar{x}_i^2}\left[V(\bar{y}_i) + R_i^2 V(\bar{x}_i) - 2R_i Cov(\bar{x}_i, \bar{y}_i)\right.$$

y $$Var(z) = \widehat{Var(z)} = \sum_{i=1}^{2} \frac{1}{\bar{x}_i^2}\left[v(\bar{y}_i) + r_i^2 v(\bar{x}_i) - 2r_i cov(\bar{x}_i, \bar{y}_i\right.$$

si en cada estrato E_i se seleccionan unidades de acuerdo con el diseño simple aleatorio.

$$var(z) = \sum_{i=1}^{2} \frac{1}{\bar{x}_i^2}\left[\frac{1-f_i}{n_i}\right]\left[\hat{S}_{y_i}^2 + r_i^2 \hat{S}_{x_i}^2 - 2r_i \hat{\rho}_i \hat{S}_{x_i} \hat{S}_{y_i}\right]$$

de aquí que:

$$var(z) = \sum_{i=1}^{2} \frac{1}{\bar{x}_i^2}\left[\frac{1}{n_i}\right]\left[\hat{S}_{y_i}^2 + r_i^2 \hat{S}_{x_i}^2 - 2r_i \hat{\rho}_i \hat{S}_{x_i} \hat{S}_{y_i}\right] + k$$

Por lo tanto, a medida que el término en base a las varianzas y covarianzas sea más grande n_i tendrá que recibir un mayor aporte de n [para minimizar Var(z)], es decir:

$$n_i^* \propto S_{yi}^2 + r_i^2 \hat{S}_{xi}^2 - 2 r_i \hat{\rho}_i \hat{S}_{xi} \hat{S}_{yi}$$

y $$n_i^* \propto \left(\frac{1}{\bar{x}_i^2}\right)$$

$$\therefore \quad n_i^* \propto \frac{S_{yi}^2 + r_i^2 \hat{S}_{xi}^2 - 2 r_i \hat{\rho}_i \hat{S}_{xi} \hat{S}_{yi}}{\bar{x}_i^2}$$

es decir, n_i^* depende -para su asignación- de la prueba de hipótesis de comparabilidad de R_1 y R_2 de los valores de estimación de R_i (Por ejemplo si: $\hat{R}_1 = r_1 \doteq 0$, y $\hat{R}_2 = r_2 \doteq 1$).

$$n_1^* \propto \frac{S_{yi}^2}{\bar{x}_i^2} \qquad n_2^* \propto \frac{\hat{S}_{y_1}^2 + \hat{S}_{xi}^2 - 2 \hat{\rho}_i \hat{S}_{xi} \hat{S}_{yi}}{\bar{x}_i^2}$$

Lo cual parece obvio pero no se percibe explícitamente en las ilustraciones anteriores.

Por otra parte, se exhibe que en estos casos se exige conocimiento(información) sobre el valor de las variables que representan cada característica(en cada estrato). Si no se dispone de muestras previas o piloto, se presentan varios cursos de acción

posibles, ¿nos quedaríamos con asignación proporcionales o iguales en cada estrato?. Este tipo de preguntas requiere resolverse en base a un marco de referencia más amplio, que sin embargo es bastante difuso.

Hasta este punto, entonces, y con el solo propósito de abrir cauces a la reflexión se quedará la exposición de un problema que presenta amplias posibilidades a la investigación.

Es claro que, además, la relación que existe entre las Encuestas Analíticas y los Diseños Experimentales(De hecho, una Encuesta Analítica por muestreo es un caso particular de los Diseños Cuasi-Experimentales y Pseudo-Experimentales). Las soluciones a las preguntas planteadas pueden orientarse a dicha vinculación.

CONCLUSIONES

a) Uno de los usos más frecuentes -o más comunes- del muestreo es el que se refiere a conteo o enumeración. Este caso, como se hace explícito en la tesis, es el caso descriptivo. Unicamente que es, entonces, un enfoque exploratorio. Cuando se requiere una mayor profundidad(en la investigación) se necesitan establecer hipótesis estadísticas para hacer evidentes los alcances del diseño de muestreo que se propone. Este punto, en mi opinión, se resalta en la tesis.

b) Un diseño de muestreo sólo puede tener sentido, en la investigación social, si corresponde primero, a un diseño de estudio subyacente, y a un diseño de encuesta. Las Encuestas por Muestreo, entonces, se desprenden de un marco teórico más amplio y requieren de un contenido de aplicación específico. Este es el resultado que pretendí sugerir constantemente en la tesis, aunque es cierto que se requeriría por completo otro proyecto de tesis para fundamentarlo como se debe; los textos de Moser y Kalton, Selltiz y Des Raj contienen elementos precisos para corroborar esta aseveración.

c) La distinción entre muestreo estratificado y el manejo del concepto de "dominios de estudio" es, para los usuarios del muestreo, a menudo - borrosa. En mi opinión, como lo revela esta tesis, los dos conceptos no son equivalentes. Aún más, se pueden incluir estratificaciones que no incluyan "dominios - de estudio". Por otra parte, se observa incongruente que se consideren "dominios de estudio" sin que haya, por lo menos, etapas de estratificación relevantes. En el caso de "subclases" la afirmación anterior parece obvia. En la situación de análisis de clases cruzadas o clases transversales se exige la inserción de estratos.

d) Para delimitar el tamaño de muestra requerido en una encuesta analítica se necesita hacer uso de planteamientos diversos a los comunes en en -- cuestas descriptivas. Esta omisión, cuando se presenta, puede ser muy costosa en el análisis estadístico posterior.

e) Además, existe una gran variedad - de dominios de estudio. Se presentaron, en general, - la clasificación y técnicas de estimación en dominios pequeños. Por otra parte, se señalaron algunas de las respectivas limitaciones y alcances. Podría, entonces,

obtenerse una conclusión un tanto decepcionante:" toda situación es casuística y no hay mejor método para cada caso". No obstante, la conclusión que presento aquí sé refiere a dos rubros diferentes en el muestreo probabilístico; la determinación del tamaño de muestra en encuestas analíticas, para obtener niveles prefijados de precisión y confianza - influye en el procedimiento de selección(de unidades muestrales) y explícitamente(vía la varianza y otros indicadores) en el - procedimiento de estimación. Así, si se considera una taxonomía de "dominios de estudio" y se plantea en - forma apropiada las hipótesis suyacentes, el tamaño - de la muestra planeado cubrirá las metas fijadas y, para cada caso en particular, se podrá revisar los efectos de cada modalidad en el Diseño de Muestreo.

f) Finalmente, se continua reflexio - nando en esta tesis sobre una inquietud importante en los cursos de Muestreo en la Facultad de Ciencias: El uso del muestreo debe contestar a un ¿porqué?, a un - ¿para qué? y a un ¿cómo?. Los trabajos sobre distribución del tamaño de muestra en estratos haciendo uso - de técnicas de Programación Matemática(por ejemplo: - Programación No lineal) son eslabones de una cadena - que busca precisar la metodología que se pueda utilizar. En este punto, el concepto de "dominios de estu-

dio" como herramienta de la investigación debe ser difundido y aclarado en la labor docente desde licenciatura para que el ejercicio práctico de los muestristas se mejore.

R E F E R E N C I A S

Cochran, G.W.(1976). Técnicas de Muestreo. Editorial Continental, México.

Cochran, G.W. y Cox, G.M.(1957). Experimental Design. John Wiley and Sons, New York.

Deming, W.E.(1968). Sample Design in Business Research. John Wiley and Sons, New York.

Flores, A.M.(1954). Métodos y Aplicaciones del Muestreo Estadístico. Sría. de Economía. Depto. de Muestreo.

Hansen, Hurwitz & Madow.(1953). Sample Survey Methods & Theory. Vol.II. John Wiley and Sons, New York.

Hess, I., Sethi, V.K. "Stratification: Apractical Investigation". J.A.S.A., 61, 74-90,(1966).

Jessen, J.R. (1978). Statistical Survey Techniques, - John Wiley and Sons, New York.

Purcell, J.N., Kish, L.(1979). "Estimation for Small - Domains". Biometrics, 35, 365-384.

Ranjar, K.S. A manual of Sampling Techniques. Heinamann Educational Books L.T.D.(1973).

Román, M.F.(1973). "Uso y mal uso de técnicas de Estratificación". México, "Curso Latinoamericano de Estadística en la Seguridad Social. CIESS, IMSS.(Memo - ria general del curso).

Selltiz, C. Jahoda, M. Métodos de Investigación en las Relaciones Sociales. Editorial Rialp Madrid.(1965).

Stuart, A. Basic Ideas of Scientific Sampling. Charles Griffin.(1976).

Sukhatme, P.V. Teoría de Encuestas y Aplicaciones. Fondo de Cultura Económica. México.

Kish, L.(1962). Teoría de Encuestas.

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

Dalenius, T., Hodges, J. "The choice of Stratification points". Skandinavis Aktuarietids. Vol. 3-4, 198-203.(1957).

Des, R. The Design of Sample Survey. John Wiley and - Sons, New York.

Evans, W.D.(1951). "On stratification and optimum - allocation". J.A.S.A. 46, 95-104.

Gosh, S.P.(1958). "A note on stratified random sampling with multiple characters". Calcutta. Statistical - Association Bulletin, 8, 81-89.

Holguín, Q.F., Hayashi, M.L. Elementos de Muestreo y - Correlación. Textos Universitarios.(1974).

Kish, L.(1961). "Efficient Allocation of a Multipurpose sample". Econometrica. 29, No.3.

Kish, L., Anderson, D.W. "Multivariate and Multipurpose Stratification". J.A.S.A. 73, 24-34, No.361.(1978).

Kish, L., Frankel, M.R.(1974). "Inference from Complex Samples". Journal of the Royal Statistical Society. B, 36, 1-37.

Moser, C.A., Kalton, G.(1972). Surveys Methods in Social Investigation. N.Y. Basic Books.2a. edición.

Oppenheim, A.N. Cuestionary Design and Attitude Mesure-- ment. Heinamann Educational Books. Londres.

Purcell, J.N., Kish, L."Postcensal estimates for local areas(or domains)". International Statistical Review. 48, 3-18.(1980)..

Selltiz, C. Métodos de Investigación en las Relaciones - Sociales. Editorial Rialp Madrid.(1965).

Sukhtame, P.V., Sukhatme, B.V.(1970). Sampling Theory of surveys with applications. IOWA State University.

Unión Panamericana. Estadística. Métodos. Instituto Interamericano de Estadística.