Tamaño de la muestra y análisis de asociación

Fernardo Cortés

I..Presentación del problema

Consideraciones de orden técnico, entre las que se cuentan la naturaleza del marco muestral, los niveles de confianza con que se desea trabajar y los errores máximos que estamos dispuestos a tolerar, permiten establecer el número de unidades que deberán componer la muestra.

Por otra parte, en toda investigación por muetreo existen restricciones presupuestarias que llegan a expresarse en un tamaño muestral máximo posible.

La comparación entre el tamaño de muestra que emerge de argumentos técnicos, con el que surge de consideraciones económicas puede dar lugar a tres situaciones: i) Las restricciones presupuestarias arrojan un tamaño de muestra mayor que el determinado por consideraciones técnicas. En esta situación el investigador puede tomar uno de dos caminos: mantener los niveles de precisión previamente establecidos y usar sólo una parte del presupuesto, o bien gastar todo el dinero disponible con lo cual se puede aumentar los niveles de confianza o disminuir los tamaños esperados de los errores muestrales. ii) El tamaño de muestra determinado técnicamente es mayor que el permitido por las restricciones presupuestarias. En este caso habría que examinar los efectos que tendrá sobre los niveles de precisión y de error el hecho de usar como tamaño de muestra el número de observaciones permitidas por las restricciones financieras. En caso de que al realizar este análisis se tenga como resultado que con el tamaño de muestra económico no se cumple las condiciones técnicas mínimas, será necesario tratar de acopiar mayor cantidad de recursos monetarios de manera que el número de observaciones muestrales se encuentre en el interior de la región técnicamente aceptable. Si estos esfuerzos resultasen infructuosos sólo quedaría abierta la alternativa de abandonar la idea de obtener información a través de procedimientos muestrales. iii) Coincidencia entre el tamaño de muestra económico y técnico. Este caso no involu-

cra ningún problema de decisión, aunque raras veces se presenta en situaciones prácticas.

Es usual que al examinarse los factores que subyacen a la determinación del tamaño de muestra se desplieguen argumentos de orden técnico y financiero análogos a los que hemos presentado. Sin embargo, rara vez se hace explícita la relación entre los objetivos de la investigación muestral y el número de observaciones que de ellos se derivan.

En efecto, gran parte de la teoría del muestreo se dedica a los problemas de estimación de parámetros poblacionales de carácter descriptivo como son las medias y variantes poblacionales. Los procedimientos de determinación de tamaño de muestra no escapan a esta limitación en la medida que usualmente responden a preguntas relativas a estimaciones de promedios o de totales. Pero, en ciencias sociales es corriente extraer muestras para analizar relaciones entre variables, las cuales, además, casi siempre son cualitativas. En este tipo de estudio rara vez interesa conocer los valores poblacionales de medidas descriptivas. El propósito central de utilizar procedimientos muestrales como instrumento de recolección de información consistiría básicamente en estudiar la presencia o ausencia, la forma y la fuerza de la relación entre variables. En consecuencia, el problema de determinación de tamaño de muestra abandona el ámbito de la estadística descriptiva de variables métricas para ubicarse dentro de la estadística de atributos.

La estadística de atributos no sólo permite estudiar el grado de relación entre pares de variables, sino también trabajar simultáneamente con el cruce de tres o más variables. La introducción de variables cualitativas en el análisis estadístico, establece una demanda creciente de observaciones, que de no satisfacerse conduce a una inadecuación en la aplicación de algunas técnicas de análisis.

Consideremos, a manera de ejemplo, que estamos interesados en apoyar empíricamente la hipótesis de una fuerte relación entre la composición orgánica del capital y el nivel de conflicto abierto que afecta al sector industrial de una sociedad. Supongamos que contamos con una serie cronológica de datos por empresas en que se ha consignado tanto la composición orgánica del capital como el número de huelgas que las ha afectado. Con esta información podríamos construir una tabla de dos por dos como puede verse en el cuadro 1.

El cruce de variables ha hecho caso omiso de la fecha a que se refieren los datos, sin embargo ésta puede ser información de importancia si la hipótesis también sostiene que la relación entre las dos variables está afectada por la ubicación estratégica de las empresas en el modelo de acumulación. * De esta manera, el sostén empírico de las ideas expuestas requiere del análisis estadístico de unas tablas como las de el cuadro 2.

* La clasificación de las empresas en la dicotomía estratégica-no estratégica hace uso de la información cronológica en la medida que la ubicación de una empresa determinada en una u otra de las dos categorías dependerá de la importancia de la rama de actividad dentro del modelo de acumulación.

CUADRO 1

COMPOSICION ORGANICA DEL CAPITAL (C. O. C.)

NUMERO DE HUELGAS	ALTA	BAJA
ALTO		
BAJO		

CUADRO 2

ESTRATEGICA (C. O. C.)

NUMERO DE HUELGAS	ALTA	BAJA
ALTO		
BAJO		

(2.a)

NO ESTRATEGICA (C. O. C.)

	ALTA	BAJA
ALTO		
BAJO		

(2.b)

El número de casillas ha pasado de cuatro (4) a ocho (8) con la introducción de la tercera variable dicotomizada. A grosso modo podríamos afirmar que la incorporación de la tercera variable dicotomizada ha multiplicado por dos los requerimientos de información: hemos pasado de 2^2 casillas a $2^2 \times 2 = 2^3$ casillas.

En consecuencia, el número de observaciones necesarias para estimar la media muestral, con ciertos márgenes de precisión y con niveles probabilísticos dados, sólo por casualidad podrá ser similar a aquel que se requiere para estimar la fuerza de la relación entre dos o más variables de atributos. Por lo tanto, no debe extrañarnos que al determinar tamaños de muestras mediante las fórmulas que normalmente nos provee el muestreo

aleatorio, éstos resulten inadecuados para cumplir los requisitos estadísticos exigidos por el análisis de asociación.

Ahora bien, el objetivo básico de este trabajo es el de estudiar las peculiaridades que surgen en el muestreo cuando nos interesa determinar el número de observaciones que nos permiten realizar el análisis de asociación al nivel de las exigencias impuestas por los objetivos de la investigación. Con este propósito en perspectiva hemos optado por presentar, en primer lugar, las ideas fundamentales que dicen relación con el cálculo de tamaño de muestra en muestreo aleatorio simple. Al mismo tiempo delineamos aquellas características del análisis de asociación que tienen injerencia directa sobre el número de observaciones necesarias para satisfacer los requerimientos impuestos por los métodos estadísticos. A continuación se propone un procedimiento para determinar el tamaño de la muestra que cumpla, por una parte, con los requerimientos técnicos que derivan del análisis estadístico de atributos y que por otra, se vincule a los requerimientos impuestos por las necesidades del análisis teórico.

Antes de entrar al corpus de este trabajo es necesario aclarar que la solución que proponemos se deriva de la aplicación de un criterio que puede o no ser objeto de discusiones. Sin embargo, en esencia, el mensaje que deseamos transmitir se refiere a la no consistencia lógica entre los métodos comunes en uso para determinar tamaños de muestras y el tipo de preguntas que emergen desde el ámbito de las ciencias sociales respecto al número de observaciones necesarias para estar en condiciones de realizar un análisis estadístico que permita el cruce simultáneo de un conjunto determinado de variables cualitativas.

II. El tamaño de muestra en muestreo aleatorio simple

Uno de los aspectos básicos que se debe considerar al aplicar un muestreo se refiere al número de unidades que conformarán la muestra. En la sección anterior hemos destacado la *importancia práctica* que tiene para cualquier investigación muestral el disponer de una idea acerca de las consecuencias (en términos de tamaño de muestra) que se derivan de los criterios técnicos. A continuación nos preocuparemos por estudiar las *vinculaciones teóricas* entre las normas de precisión y el tamaño de la muestra.

El primer elemento que se asocia con el tamaño de la muestra (simbolizado por n) es la discrepancia máxima que el investigador está dispuesto a aceptar (denotada por d) entre la media muestral y el promedio de la población. Si bien aún no se conoce la media muestra ($\bar{x}$), porque la determinación del tamaño de muestra es previo al muestreo mismo, ni

tampoco se sabe el valor de la media poblacional (μ), ello no impide que el investigador pueda fijar de antemano el error máximo que está dispuesto a admitir, o en otros términos, la discrepancia máxima admisible entre $\bar{x}$ y μ.

Hay que destacar que d es el elemento que introduce el criterio básico que guía el cálculo del tamaño de la muestra: estimar la media de la población μ, con un nivel de precisión prefijado. En otros términos, *se trata de obtener un número de unidades muestrales tal que permita cumplir con el criterio de que la media muestral no se desvíe de la poblacional en más que una cierta cantidad.*

A medida que es mayor el valor del error máximo admisible (es decir, en cuanto menor es la precisión) que se exige en la investigación menor será el tamaño de la muestra y a mayor precisión mayor tamaño de muestra. En otros términos la relación entre n y d es inversa.

El segundo elemento que se asocia con n es el coeficiente de confianza (simbolizado por t). El hecho que el error máximo admisible se defina como $d = \bar{x} - \mu$, implica que será imposible garantizar que en *una muestra específica* se alcance un valor igual o menor que el fijado previamente por el investigador. El tamaño que asuma d dependerá directamente del valor de la variable aleatoria $\bar{x}$, es decir, de la muestra particular que haya sido seleccionada. En consecuencia, estaremos incapacitados para realizar previsiones puntuales referidas a d, pero ello no impide que tomemos precauciones para garantizar que un porcentaje alto de las muestras posibles de ser seleccionadas cumplan con la condición de generar errores muestrales menores o iguales al fijado por el investigador.

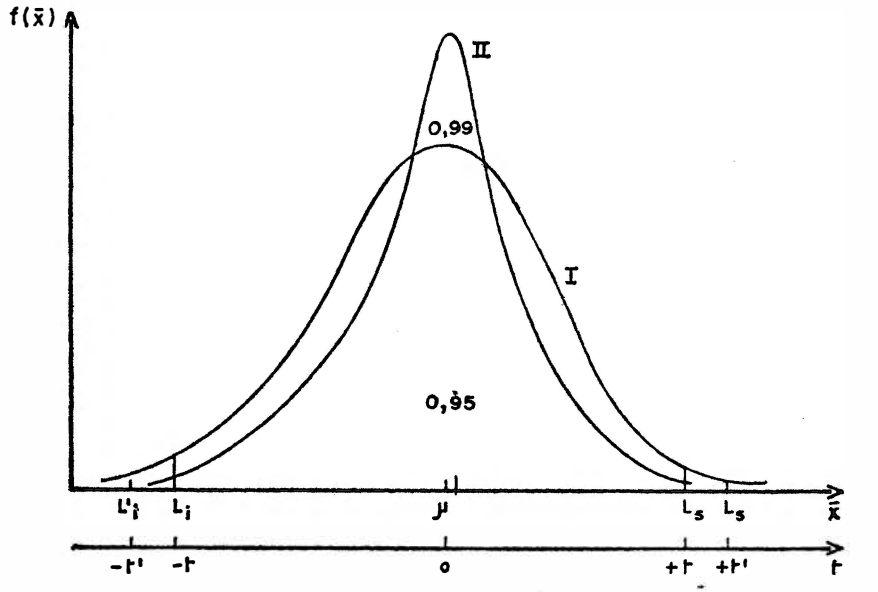
Ahora bien, en el gráfico hemos representado la distribución de frecuencia de las medias muestrales e incluido el criterio que plantea que el error de muestreo no debe ser superior a d. La aplicación de este criterio genera dos límites, uno superior (L_s) y otro inferior (L_i) que definen un conjunto de valores de medias muestrales que cumplen con la condición de precisión (que la discrepancia con μ sea menos que d).

Como las medias muestrales tienden a distribuirse conforme una distribución t de Student,* la proporción de muestras que entregarán como resultados medias aritméticas en el intervalo definido por los límites superior e inferior será igual al área que le corresponda en la curva de probabilidades. En consecuencia esta área puede ser interpretada como el nivel de confianza que se utiliza para tomar la muestra. En el gráfico hemos supuesto (curva I) que el nivel de confianza es igual a 0.95. Es decir, que esperamos que de cada 100 muestras 95 entreguen como resultado

* Según el teorema central del límite las medias muestrales tienden a distribuirse normalmente en la medida que n tiende a infinito. Cuando se desconoce la varianza de la población (σ^2) y se estima a través de la varianza muestral entonces la variable aleatoria $\bar{x}$ sigue una distribución t de Student. Al respecto ver por ejemplo: Lowell Wine *Statistics for Scientists and Engineers.* Prentice Hall, 1964, p. 250.

valores para $\bar{x}$ en el intervalo definido por L_i y L_s y que sólo en 5 de cada 100 muestras el valor de la media muestral escapará a estos límites.

GRAFICO 1



En el gráfico hemos agregado un eje horizontal *t*, paralelo al eje de abscisa $\bar{x}$ y que se relaciona con éste a través de:

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

donde todos los términos han sido previamente definidos, excepto *s*, que simboliza la desviación típica muestral. ** La variable *t* corresponde a una estandarización de $\bar{x}$ y la fórmula que la define establece una regla de correspondencia entre los ejes horizontales que hemos incluido en el gráfico. A través de *t* es posible acceder a las tablas de la distribución de Student y determinar la probabilidad que corresponde al intervalo.

Un aumento del nivel de confianza, por ejemplo de 0.95 a 0.99, tendrá dos efectos posibles: i) Un incremento en el valor de *d*, siempre que no hubiese alteración en la curva de probabilidades (no cambiará su posición ni su dispersión) el que se reflejará en un crecimiento del coeficiente de confianza *t*. En el gráfico se puede apreciar que el nuevo error máximo admisible d', queda definido por los límites L'_i y L'_s, lo que se traduce en

** La varianza muestral con denominador (n - 1) es el estimador no sesgado de la varianza poblacional.

mayores valores en la escala del coeficiente de confianza *t*. Por lo tanto, a mayores niveles de confianza corresponderán mayores coeficientes de confianza. ii) Un cambio en la dispersión de la curva si decidimos mantener constante el error máximo admisible. En efecto, si se decide mantener *d*, la única alternativa de incluir una mayor porción de área bajo la curva consistirá en disminuir su variabilidad. La dispersión de la distribución de probabilidades (que responde a la fórmula $s/\sqrt{n}$) se puede disminuir a través de un aumento del tamaño de la muestra o bien bajando el valor de la dispersión representada por *s*. Esta última alternativa queda fuera de las posibilidades de manejo del investigador ya que se relaciona con la dispersión de la población, la que es un dato: en general a mayor dispersión de la población corresponderá un mayor valor de *s*. En consecuencia, el único camino viable para aumentar el nivel de confianza será el de tomar muestras con un número mayor de unidades, lo que nos permite concluir que dado el error máximo admisible, a mayores niveles de confianza corresponderán mayores tamaños de muestras.

Por otra parte habrá una relación directa entre la dispersión de la distribución de frecuencias y el tamaño de la muestra. Independientemente del número de unidades que compongan la población (N), mientras mayor sea la concentración de la variable, menor será el *n* necesario para estimar μ con un *d* fijo y un nivel de confianza dado. Es evidente que si el tamaño de la población A_1 es sustancialmente menor que el de la población A_2, pero a diferencia de A_1, A_2 presenta escasa dispersión, seguramente el tamaño de muestra requerido será menor (manteniendo los demás factores constantes). En el caso límite en que una población tenga varianza cero, bastará con una muestra de tamaño uno para estimar la medida de la población. Tal sería el caso, por ejemplo, en que se intenta estimar la estatura promedio de una población en la que todos los miembros miden exactamente 1.72 mts. o bien, en que se trata de estimar la tasa de desocupación de una población plenamente ocupada.

En general, el tamaño de la población tendrá una relación directa con el número de unidades que deberán componer la muestra. Si mantenemos constantes los restantes elementos que juegan en la determinación de *n* (d, t, s^2), tendremos que a mayor *N* deberá corresponder también un *n* mayor.

Hemos visto que el tamaño de muestra que surge de argumentos únicamente técnicos se relaciona directamente con: i) el nivel de confianza, ii) la dispersión de la distribución y iii) el tamaño de la población. E inversamente con el error máximo admisible. El *n* que deberá tomarse en una situación concreta surgirá de la consideración simultánea de todos estos factores, los cuales se conjugan en la siguiente expresión matemática:

$$(2) \qquad n = \frac{n_0}{1 + \frac{n_0}{N}}$$

donde:

$$(3) \qquad n_0 = \frac{t^2 \, s^2}{d^2}$$

En la medida que el tamaño de la población tiende a infinito la igualdad (2) tiende a confundirse con la (3). Es decir $\lim_{N \to \infty} n = n_0$ en consecuencia (3) nos entrega una fórmula para calcular el tamaño de la muestra cuando el tamaño de la población puede ser considerado infinito. Desde el punto de vista del muestreo aplicado algunos autores recomiendan utilizar (3) siempre que la fracción de muestreo (n/N) sea menor que 0.10 y otros en aquéllos casos en que sea inferior a 0.05. La igualdad (2) deberá utilizarse sólo si es adecuado suponer que la población es finita, concepto que se define como complementario al de población infinita.

Hasta este punto hemos considerado los elementos técnicos que juegan en la determinación del tamaño de muestra y la forma particular como se combinan en una expresión matemática. Pero lo más importante para los propósitos de este trabajo es dejar claramente establecido que el objetivo central de esta fórmula tiene que ver con la estimación de la medida poblacional. Es una respuesta a la pregunta ¿De qué tamaño debe ser la muestra para estimar la media de la población (μ) con un error máximo admisible no mayor que *d* y con un nivel de confianza de *a* por ciento? La respuesta a esta pregunta, sólo por casualidad tenderá a confundirse con la de la pregunta ¿Cuál es el número de observaciones muestrales que se requieren para realizar un análisis de contingencia en que se cruzan simultáneamente un cierto número de variables cualitativas (por ejemplo cuatro variables) constituidas por algunas categorías (por ejemplo, dos dicotómicas, una tricotómica y una tetracotómica)?

Pero antes de abandonar esta sección nos referiremos a un par de temas adicionales que si bien tienen importancia desde el punto de vista del muestreo aplicado sólo tienen una vinculación tangencial con el corazón de este trabajo. Han sido incluidos debido a que más adelante necesitaremos contrapuntear los pasos previos a la recolección de la muestra que requiere el procedimiento tradicional con aquél que proponemos en la sección IV.

El cálculo del tamaño de muestra, según las fórmulas que hemos presentado, requiere del conocimiento de la varianza muestral en circunstancias en las que aún no se ha procedido a tomar la muestra. Para resolver estas *impasse* se puede recurrir a i) la información que entregan sobre varianzas otras muestras que hayan trabajado la variable; ii) información censal o iii) la aplicación de una muestra de iluminación o muestra piloto.

Respecto a la información censal sobre varianza, resulta obvio que el investigador no espera encontrar la varianza que le interesa, porque, en ese caso lo más probable es que también se conozca la media aritmética y por consiguiente no necesitaría tomar una muestra. Lo que debe buscar

es información que le permita estimar la varianza. Por ejemplo, puede buscar la varianza de una variable que tenga una relación conocida o posible de estimar con la que interesa o si la información está desfasada en el tiempo, postular hipótesis respecto al desarrollo temporal de la variable, o de la variable relacionada.

La alternativa de la muestra piloto es la más utilizada, en investigaciones aplicadas debido al escaso desarrollo de los métodos para utilizar la información censal y a la precaria accesibilidad a datos censales desagregados. La muestra de iluminación tiene como objetivos básicos estimar la varianza para alimentar la fórmula de tamaño de muestra y probar los cuestionarios de la encuesta.

Normalmente, no se plantea la posibilidad de usar la información de la muestra de iluminación, en la muestra definitiva. Sin embargo bajo ciertas condiciones esto parece perfectamente posible. Si la variable tiempo no juega un papel determinante en las características de las distribuciones de las variables bajo estudio, si el tiempo que media entre la obtención de la muestra piloto y la definitiva no es extremadamente prolongado y el cuestionario no presenta deficiencias graves en las variables más importantes, se podrían usar las observaciones de la muestra de iluminación en la muestra final. Esta misma estrategia, se puede seguir en el caso en que el tiempo afecte sustancialmente las distribuciones de las variables, pero el período entre la muestra de iluminación y la definitiva sea lo suficientemente breve como para que el impacto no sea significativo. Debe entenderse que en este caso, se está suponiendo que el cuestionario no presenta mayores problemas.

El tamaño de muestra que entregan las fórmulas se refiere a sólo una variable, en circunstancias que rara vez se toma una muestra para conocer las características de la distribución de una variable. Lo más usual es diseñar una muestra para investigar un conjunto de variables.

Para alimentar la fórmula, debe decidirse respecto a qué variable se va a calcular el tamaño de la muestra. Ahora bien, podríamos recurrir al criterio de calcularlo para aquella que sea más exigente en términos de unidades muestrales. En otros términos para aquella variable que necesite un mayor nivel de confianza, una estimación más precisa y que tenga una mayor varianza. En este criterio, se cumplen los requisitos técnicos de las otras variables, con mayor rigurosidad que las planteadas por el investigador. También se usa a veces un criterio promedio. Pero, antes de aplicar algunos de los criterios reseñados, es conveniente clasificar las variables que se van a investigar, en aquéllas que son esenciales para la investigación y aquéllas que no lo son. Una vez hecha esta decisión podremos operar según los criterios expuestos.

III. El proceder del análisis de asociación

Así como en la seccion anterior no pretendimos agotar el tema de determinación del tamaño de muestra en muestreo aleatorio simple, en ésta no intentaremos llevar a cabo una exposición detallada del análisis de asociación en que se cruzan simultáneamente varias variables. Sólo mostraremos la estructura de tablas que se generan en este tipo de estudios por considerarse que este aspecto de la aplicación del análisis de asociación es el que se entronca directamente con el argumento de este escrito.

El estudio del grado de relación entre variables cualitativas normalmente * se aborda a través de la clasificación en tablas cruzadas de la información entregada por censos o encuestas. Una vez que disponemos de la tabla bidimensional de frecuencias, se puede continuar el análisis examinando la relación original en el interior de distintos subcolectivos, definidos ya sea por una o por el cruce de varias variables.

En esta perspectiva se abren dos alternativas al investigador. Una consistiría en examinar el grado de asociación entre dos variables en subcolectivos definidos por una tercera. El otro camino lo constituiría el examen de la relación en subcolectivos definidos por conjuntos de variables. Si bien esta distinción parece ser básicamente formal, pierde tal connotación en la medida que enfocamos ésta distinción desde el punto de vista de control de variables.

Al estudiar, a través de una tabla cruzada, la presencia o ausencia y la fuerza de la relación entre dos variables podemos llegar a tener una visión distorsionada de ella. En efecto, el nivel de asociación que percibimos puede estar condicionado por una tercera variable y de esta manera se nos puede aparecer una relación, aunque ella sólo sea aparente. Paul Lazarsfeld ** plantea a título de ejemplo una relación entre el número de cigüeñas y la cantidad de nacimientos, la cual desaparece una vez que se controla por la urbanización. Se presenta a los ojos del investigador debido a que en las zonas rurales donde hay una mayor cantidad de cigüeñas la natalidad tiende a ser más elevada que en las ciudades donde hay simultáneamente menos nacimientos y cigüeñas. Al definir dos subcolectivos, uno rural y otro urbano y examinar la asociación en el interior de cada uno de ellos se controla la variable urbanización y consecuentemente la relación aparente desaparece en el interior de cada contexto.

El análisis de la asociación entre un par de variables dentro de subcolectivos definidos por una tercera, sólo permite el control estadístico de una variable cada vez. Sin embargo, puede acontecer que el nexo que

* Decimos que "normalmente" porque el análisis de regresión también permite incorporar variables cualitativas, aunque este uso se encuentra muy poco difundido. Ver por ejemplo: Johnston J. *Econometric Methods*. John Wiley, 1972. p. 176.

**Lazarsfeld Paul, "La interpretación de las relaciones estadísticas como propiedad de investigación" en: Boudon y Lazarsfeld, *Metodología de las Ciencias Sociales*, Laia, 1974, p. 29.

las une se encuentre afectado por más de una variable. En este caso el control estadístico requerirá que se definan subcolectivos a través del cruce de ellas para luego proceder a caracterizar la asociación. En este caso la estrategia del control de sólo una tercera variable cada vez resultaría inadecuado.

Ahora bien, el control simultáneo de varias variables origina un número de tablas igual al producto del número de categorías contenidas en cada variable. Así por ejemplo, si se estudia la relación entre dos variables en un subcolectivo definido por dos variables control, una dicotómica y otra tricotómica, entonces el número de tablas será igual a $2 \times 3 = 6$, es decir se trata de estudiar la asociación en seis subcolectivos.

En función de los desarrollos que presentaremos más adelante nos interesa distinguir los *niveles* en que estudiaremos la relación, los cuales se definen en términos del *número de variables control* utilizadas: si se usa una variable control diremos que las tablas se encuentran en el nivel 1, si son dos entonces las tablas se ubicarán en el nivel 2 y así sucesivamente; se dirá que la tabla original está en el nivel cero.

En la medida que aumenta el nivel en el cual deseamos estudiar los vínculos entre las variables, la cantidad de información por tabla disminuye en promedio. Un fenómeno similar ocurre dentro de un mismo nivel si se aumenta el número de categorías por variable. En otros términos, la cantidad de información por tabla tenderá a disminuir en la medida que mayor sea el número de categorías involucradas en las variables de la relación original y mayor sea el producto de las categorías que conforman las variables control. Este mismo hecho pero mirado al revés, nos dice que dada una cierta cantidad de información y el número de categorías de las variables, habrá limitaciones al número de variables que se podrá considerar simultáneamente, es decir, habrá un límite al número máximo de variables que se podrán controlar simultáneamente el cual en ocasiones puede llegar a ser extremadamente reducido (no más de dos variables control). Esta restricción estadística puede llegar a ser tan fuerte que impida o restrinja el trabajo que había sido diseñado tomando en cuenta únicamente elementos teóricos.

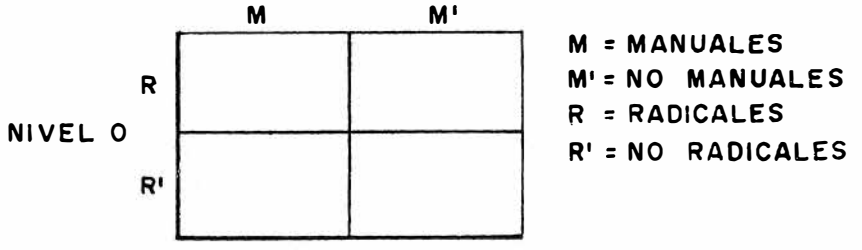
En lugar de suponer que la cantidad de información estadística está dada, podríamos partir desde el dominio de la teoría e intentar establecer de antemano la cantidad de niveles que resultarían adecuados, así como el número de categorías de cada variable, ambas, tanto las categorías como las variables, consistentes con las preguntas teóricas que han originado la investigación, y a continuación preguntarse por el número de observaciones que se requieren para llevar a cabo el trabajo. Como se puede apreciar la interrogante que se ha levantado es bastante diferente a la planteada por la teoría estadística del muestreo.

Para una mejor comprensión de las consideraciones que hemos expuesto procederemos a desarrollar un ejemplo. Supóngase que un investigador

está interesado en estudiar el comportamiento político de los trabajadores en función de sus inserciones en el aparato productivo. *

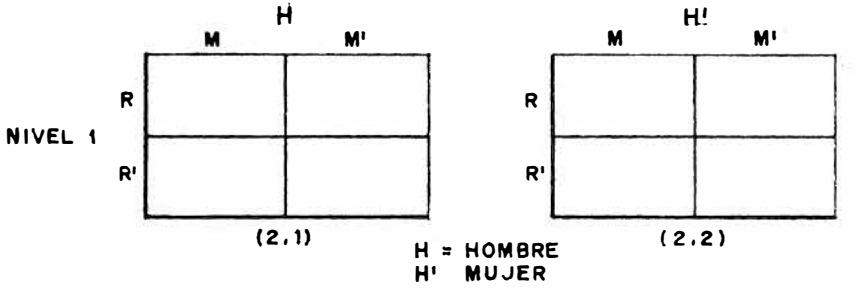
Para ello, define la variable radicalismo político y la dicotomiza en radicales y no radicales, a su vez, a los trabajadores los clasifica en manuales y no manuales. Al cruzar ambas variables se genera una tabla con la estructura de la siguiente.

TABLA 1



Una vez que se ha analizado la tabla anterior, el investigador desea estudiar la relación en los subcolectivos de hombres y mujeres, para lo cual genera las dos tablas cuya estructura se presenta a continuación:

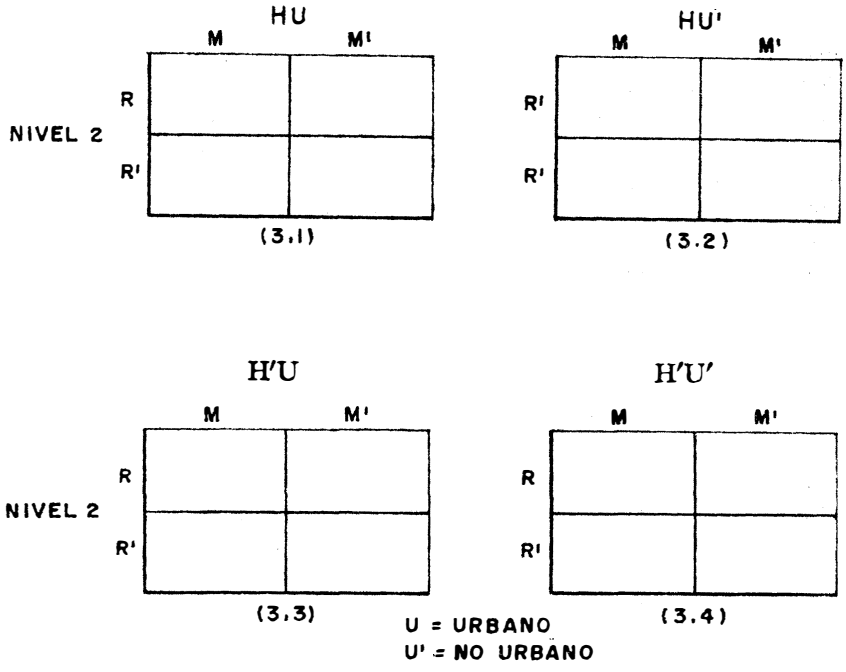
TABLA 2



Por último, interesa estudiar la relación en los subcolectivos hombre urbano, hombre rural; mujer urbana y mujer rural. Estos subcolectivos se obtienen, abriendo las tablas anteriores, en dos cada una, generándose de este modo cuatro tablas.

* En éste artículo, se supone que el número de variables y sus características se determinan en función de la teoría sustantiva que el investigador está manejando. En la redacción de esta sección sólo se están considerando las consecuencias del discurso teórico.

TABLA 3



Supongamos que los intereses teóricos del investigador lo llevan a plantear el análisis estadístico a través de la estructura de tablas que hemos presentado. Se trata entonces de realizar un análisis que llega hasta el segundo nivel. Según la definición de nivel que hemos entregado, la primera tabla será de nivel cero porque no hay variables control. Las dos tablas que vienen a continuación se encuentran en el nivel 1 porque hay una variable control, el sexo. Las cuatro tablas restantes se ubican en el nivel 2 porque los subcolectivos se definen por el cruce de las variables sexo y urbanización.

El examen de las tablas puede llevar a concluir que el análisis de asociación consiste en la simple estratificación de una población en que se procede al estudio de la relación entre atributos en el interior de cada estrato. De aquí no sería demasiado difícil extender un poco más el razonamiento para concluir entonces que el diseño muestral apropiado para este tipo de problemas sería el aleatorio estratificado. Sin embargo, no debe olvidarse que el diseño tiene como propósito central hacer más eficiente el proceso de estimación de parámetros poblacionales descriptivos y que al científico social le interesa fundamentalmente disponer de un número de casos que le permita llevar a cabo los análisis que se derivan de su pensamiento teórico.

IV. Tamaño de muestra y análisis de asociación

Así como en la aplicación usual del muestreo aleatorio simple interesa llegar a tener una idea aproximada respecto al número de observaciones que son necesarias para estimar la media de la población con ciertos niveles de precisión y confianza, en el ámbito del análisis de asociación también debemos preocuparnos por determinar un tamaño de muestra tal que permita realizar la investigación empírica en concordancia con nuestras preocupaciones teóricas.

La pregunta que nos hemos propuesto sólo podrá tener una respuesta si el discurso teórico se encuentra lo suficientemente desarrollado como para ayudar a construir los indicadores empíricos de las variables teóricas así como sus categorías. En otros términos, la condición mínima que posibilita la elaboración de una respuesta a la pregunta que nos hemos formulado será la de un desarrollo teórico que permita establecer un plan de cruces de variables.

En esta sección nos proponemos como problema indagar respecto a la posibilidad de levantar una respuesta a la pregunta señalada. Trabajaremos con el supuesto que el plan de cruces es conocido, esto quiere decir que dejamos fuera de nuestras consideraciones el examen de los vínculos entre las ideas teóricas y su traducción en un conjunto de tablas, definidas por el cruce simultáneo de variables cualitativas. Las ideas centrales serán desarrolladas, en un primer momento, sobre la base de las tablas que hemos presentado en la sección anterior y a continuación se entregará una generalización de ellas. Nos preocuparemos entonces, por indagar acerca de los elementos determinantes del tamaño de muestra en análisis de asociación. Para ello partiremos del cruce de cuatro variables dicotómicas.

El criterio que usaremos para calcular el tamaño de muestra se refiere a una de las restricciones que impone la teoría estadística para aplicar la prueba ji-cuadrada (χ^2) de independencia estadística, la cual establece que se tendrá una buena aproximación del estadígrafo χ^2 discreto a la curva continua de probabilidades χ^2, cuando todas y cada una de las frecuencias esperadas sean por lo menos iguales a cinco. *

La información que se presenta en el gráfico 2 (p. 1395) pretende recoger los elementos más relevantes que intervienen en la determinación del tamaño de muestra cuando se utiliza el criterio que las frecuencias esperadas deben ser mayores o iguales que cinco. Cada rectángulo se refiere al nivel (número de variables de control) que, en el ejemplo que estamos utilizando, alcanza el valor máximo dos. Dentro de cada uno de ellos hemos incluido el número de observaciones por tablas, donde el subíndice nos permite identificar al nivel que nos referimos, de este modo n_0 representa al

* Ver por ejemplo, Paul Hoel *Introduction to Mathematical Statistics*, John Wiley, 1962, p. 247.

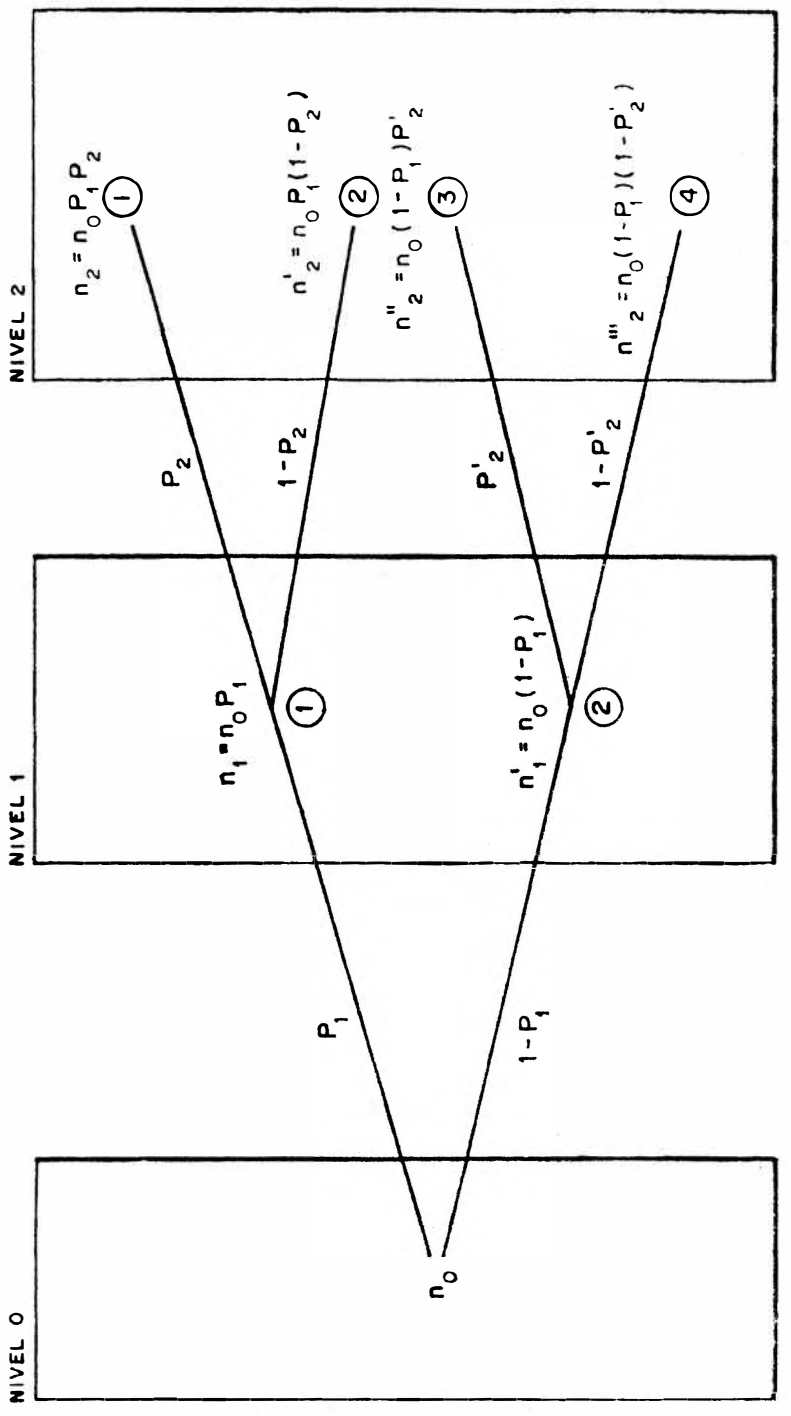

NIVEL 2
NIVEL 1
NIVEL 0
$n_2 = n_0 P_1 P_2$
$n'_2 = n_0 P_1 (1-P_2)$
$n''_2 = n_0 (1-P_1) P'_2$
$n'''_2 = n_0 (1-P_1)(1-P'_2)$
P_2
$1-P_2$
P'_2
$1-P'_2$
$n_1 = n_0 P_1$
$n'_1 = n_0 (1-P_1)$
P_1
$1-P_1$
n_0
1
2
3
4

GRAFICO 2

número de observaciones en la tabla original, mientras que n_1 y n'_1 simbolizan al número total de unidades muestrales que componen la tabla de hombres y mujeres respectivamente; n_2, n'_2, n''_2, n'''_2 denotan el número de casos en las tablas de nivel dos para los hombres urbanos, hombres no urbanos, mujeres urbanas y mujeres no urbanas respectivamente. Los rectángulos están unidos por unas líneas que llamaremos ramas y que se encuentran identificadas por un número dentro de círculo, donde además hemos anotado la proporción de observaciones de una rama de nivel inferior que pasan a formar parte de una tabla de nivel superior.

Así por ejemplo, P_1 simboliza la proporción de observaciones de nivel cero que pasarán a formar parte de la tabla de hombres de nivel 1. $(1\text{-}P_1)$, la proporción complementaria, es decir, la que formará la tabla de mujeres. Los productos $n_0\ P_1$ y $n_0\ (1\text{-}P_1)$, entregan como resultado el número de casos que esperamos encontrar en la tabla de hombres y mujeres respectivamente. Supongamos que $P_1 = 0.6$; luego $(1\text{-}P_1) = 0.4$; si en la tabla original, en que se ha cruzado la inserción ocupacional con radicalismo político contamos con 200 observaciones $(n_0 = 200)$, entonces 120 de ellas (200×0.6) pasarán a la tabla en que se ha controlado por la categoría hombres y las 80 (200×0.4) restantes a la tabla de mujeres.

Entre los niveles 1 y 2 encontramos dos probabilidades en lugar de una y sus correspondientes proporciones complementarias, ellas son la probabilidad condicional P_2 que se refiere a la proporción de urbanos entre los hombres y su complemento $(1 - P_2)$ a los no urbanos entre los hombres. De manera análoga se define P'_2 y $(1 - P'_2)$, pero esta vez para las mujeres. Siguiendo con nuestro ejemplo, supongamos que $P_2 = 0.4$ y que $P'_2 = 0.5$; luego, de los 120 hombres que teníamos en la tabla de nivel 1;48 (120×0.4) pasarán a la tabla para los hombres urbanos y 72 (120×0.6) irán a conformar la tabla en que se han clasificado los hombres no urbanos. Las 80 mujeres se distribuirán por partes iguales $(40 = 80 \times 0.5)$ entre las clasificaciones para las mujeres urbanas y rurales.

Las probabilidades P que hemos asociado a cada rama nos permiten separar del total de observaciones cuántas deben ir a cada cuadro, es decir, nos sirven para seguir la mecánica que gobierna el paso de información estadística de un nivel a otro, por ello hemos optado por denominarlas *probabilidades de distribución*. El subíndice que hemos asociado a cada P se refiere al nivel de destino, de este modo P_1 nos indica que se aplica al nivel cero o nivel de origen, para distribuir las observaciones en el nivel 1, en tanto que las probabilidades P_2 simbolizan las proporciones que aplicadas sobre los totales del nivel 1 permiten separar los casos que formarán las distintas tablas del nivel 2.

La simbología que hemos definido hasta este momento se puede resumir en:

1] Tamaño de muestra para el nivel:

a) 0 n_0

b) 1 n_1, para la rama 1 y n'_1 para la rama 2

c) 2 : n_2, n'_2, n''_2 y n'''_2 para las ramas 1, 2, 3 y 4 que unen los niveles 1 y 2, respectivamente.

2] Probabilidades de distribución para el nivel de destino:

a) 1 P_1 para la rama 1 y $(1 - P_1)$ para la rama 2.

b) 2 P_2 y $(1 - P_2)$ para las ramas 1 y 2, respectivamente, y para las ramas 3 y 4 usamos los símbolos P'_2 y $(1 - P'_2)$.

Aun cuando hemos agotado los componentes del gráfico 2, no hemos incorporado todavía todos los elementos que influyen en el tamaño de la muestra. Como el criterio que proponemos se refiere a la frecuencia que se espera en condiciones de independencia estadística, se hace necesario incorporar las frecuencias marginales, puesto que las frecuencias esperadas igualan al producto de las probabilidades marginales multiplicado por el total de observaciones de la tabla. En consecuencia el *tamaño de muestra máximo* requerido por un cruce particular resultará de considerar el producto mínimo de las probabilidades marginales. Mientras menor sea este producto, mayor será el número de observaciones que debería haber en la tabla de modo que se cumpla con la restricción. Obviamente, el producto mínimo resulta de la multiplicación de dos proporciones mínimas. Por lo tanto, sólo nos interesa distinguir las probabilidades menores, que son las que originan el tercer conjunto de símbolos que debemos distinguir por niveles.

3] Proporciones marginales mínimas para el nivel:

a) 0; en las líneas $r^0_{.\min}$ y en las columnas $r^0_{.\min}$

b) 1; rama 1. En las líneas $r^1_{.\min,1}$ y en las columnas $r^1_{\min.,2}$

1; rama 2. En las líneas $r^1_{.\min,2}$ y en las columnas $r^1_{\min.,2}$

c) 2; rama 1. En las líneas $r^2_{.\min,1}$ y en las columnas $r^2_{\min.1}$

2; rama 2. En las líneas $r^2_{.min,2}$ y en las columnas $r^2_{min.,2}$

2; rama 3. En las líneas $r^2_{.min,3}$ y en las columnas $r^2_{min.,3}$

2; rama 4. En las líneas $r^2_{.min,4}$ y en las columnas $r^2_{min.,4}$

Denominaremos casilla crítica de una tabla a aquélla que se encuentra en la intersección de la línea y la columna que tienen menores proporciones marginales.

Al aplicar el criterio de las frecuencias esperadas a su nivel menos exigente, es decir, que sean exactamente iguales a cinco tendremos:

$$4] \quad n_0 \, r^0_{min.} \, r^0_{.min} = 5$$

Despejando n_0:

$$5] \quad n_0 = \frac{5}{r^0_{min.} \, r^0_{min}}$$

De la misma manera se puede determinar la fórmula para n_1 y n_2 en los niveles 1 y 2.

$$6] \quad n_1 = \frac{5}{r^1_{min.,1} \, r^1_{.min,1}} \quad y \quad n'_1 = \frac{5}{r^1_{min.,2} \, r^1_{.min,2}}$$

$$7] \quad n_2 = \frac{5}{r^2_{min.,1} \, r^2_{.min,1}} \; ; \quad n'_2 = \frac{5}{r^2_{min.,2} \, r^2_{.min,2}} \; ;$$

$$n''_2 = \frac{5}{r^2_{min.,3} \, r^2_{.min,3}} \quad y \quad n'''_2 = \frac{5}{r^2_{min.,4} \, r^2_{.min,4}}$$

Sobre la base de estas igualdades podemos calcular los tamaños de muestra requeridos para cada uno de los tres niveles. Sin embargo, lo que interesa es saber cuál debe ser el número total de observaciones que se debe tomar de modo que se respete el criterio de las frecuencias en los tres niveles. En otros términos, hay que determinar un n_0, tal que las frecuencias esperadas, en los tres niveles y en cualquier tabla sean siempre mayor o igual a cinco.

Este problema se puede resolver usando las relaciones:

$$8] \quad n_1 = n_0 P_1 \quad y \quad n'_1 = n_0 (1 - P_1)$$

$$9] \quad n_2 = n_1 P_2 \quad ; \quad n'_2 = n_1 (1 - P_2) \quad ;$$

$$n''_2 = n'_1 P'_2 \text{ y } n'''_2 = n'_1 \ (1 - P'_2)$$

reemplazando las ecuaciones 8 en las de la 9.

$$10] \quad n_2 = n_0 P_1 P_2 \quad ; \quad n'_2 = n_0 P_1 (1 - P_2) \quad ;$$

$$n''_2 = n_0 (1 - P_1) P'_2 \quad y \quad n'''_2 = n_0 (1 - P_1)(1 - P'_2)$$

El tamaño de muestra necesario para cumplir las restricciones al nivel 0 está dado por la ecuación 5]. Este conjunto de ecuaciones se encuentran representadas en el interior de los rectángulos del gráfico 2 y expresan los tamaños de muestras para las tablas de los distintos niveles en función de número de observaciones de la tabla original (n_0) y de las correspondientes probabilidades de distribución.

Los tamaños de muestras en el primer nivel n_1 y n'_1 se obtienen reemplazando las ecuaciones 8, en las correspondientes igualdades 6.

$$11] \quad n_0 = \frac{5}{P_1 \ r^1_{\text{min.},1} \ r^1_{.\text{min},1}} \quad y \quad n_0 = \frac{5}{(1 - P_1) \ r^1_{\text{min.},2} \ r^1_{.\text{min},2}}$$

Las fórmulas para el tamaño de muestra en el segundo nivel se obtienen al reemplazar las ecuaciones 10 en las igualdades 7.

$$12] \quad n_0 = \frac{5}{P_1 \ P_2 \ r^3_{\text{min.},1} \ r^2_{.\text{min},1}} \quad ;$$

$$n_0 = \frac{5}{P_1 (1 - P_2) \ r^2_{\text{min.},2} \ r^2_{.\text{min},2}} \quad ;$$

$$n_0 = \frac{5}{(1 - P_1) P'_2 \ r^2_{\text{min.},3} \ r^3_{.\text{min},3}} \quad y$$

$$n_0 = \frac{5}{(1 - P_1)(1 - P_2')\ r^2_{\text{min.},4}\ r^2_{.\text{min},4}}$$

Para un análisis de asociación a cuatro variables dicotomizadas, se dispone de siete fórmulas. Es decir, se tendrá tantos tamaños de muestras como tablas haya definido el investigador. Esta situación, es similar al cálculo de tamaño de muestra tradicional, cuando se trabaja con más de una variable. En dicho caso, se tienen tantos tamaños de muestras, como variables distintas haya.

Dada esta situación, interesa determinar qué tamaño de muestra se va a utilizar en definitiva. Obviamente deberá utilizarse el mayor para así estar en condiciones de cumplir la restricción en todas y cada una de las tablas. Para ello no es necesario hacer todos los cálculos sino que bastará con determinar el tamaño de la muestra para aquella tabla que entregue el menor denominador. En términos del gráfico número 2 esto significa determinar la rama crítica.

El procedimiento para hacerlo consiste en:

1] Ubicarse al último nivel, en este caso el nivel 2.

2] Determinar las celdas críticas, definidas por el producto mínimo de las frecuencias marginales correspondientes.

3] Obtener la rama crítica, para lo cual es necesario determinar el producto de las proporciones marginales mínimas de la tabla por todas las probabilidades de distribución que se encuentren en la rama.

4] Se divide el valor criterio asignado a la frecuencia esperada por aquél que ha resultado de aplicar el procedimiento descrito en el punto anterior.

Al seguir los cuatro pasos señalados se determina el tamaño de muestra al nivel 0 que cumpla en todas las tablas con el criterio impuesto.

V. Un ejemplo numérico

En esta seccion se desarrolla un ejercicio numérico en que se aplica el procedimiento recién señalado. El ejemplo está basado en la clasificación

que se ha venido trabajando hasta el momento, es decir, cuatro variables dicotomizadas, en que dos de ellas se usan como variables control simultáneas.

Supóngase que se dispone de la siguiente información:

TABLA 4

	M	M'	
R			0,60
R'			0,40
	0,70	0,30	1,00

TABLA 5

$P_1 = 0,20$

H

	M	M'	
R			0,30
R'			0,70
	0,10	0,90	1,00

(5.1)

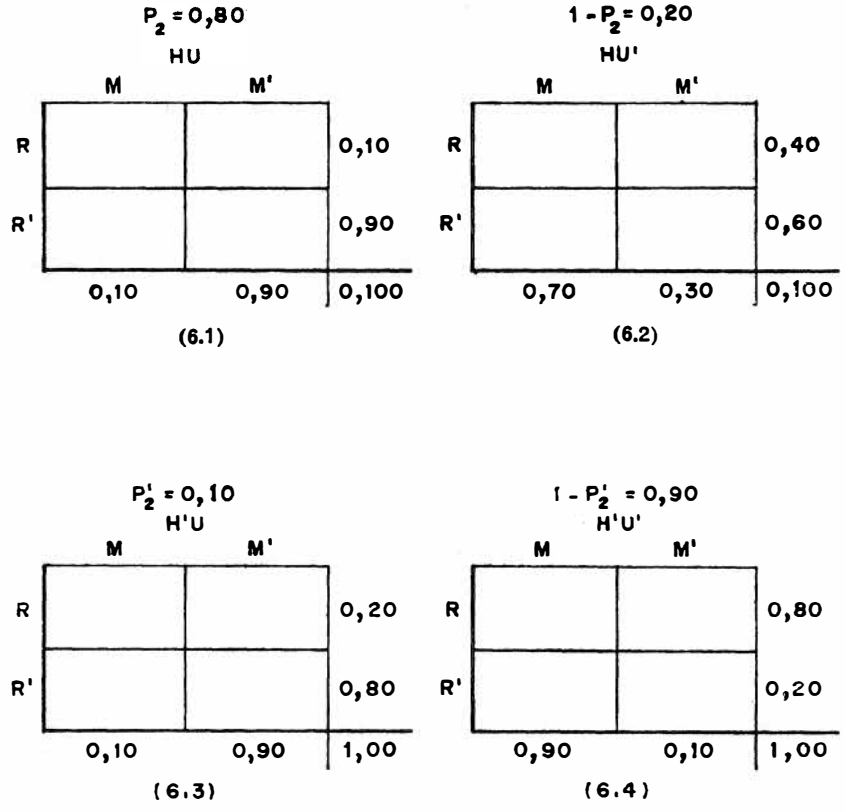
$1 - P_1 = 0,80$

H

	M	M'	
R			0,80
R'			0,20
	0,80	0,20	1,00

(5.2)

TABLAS 6



Si se está interesado en determinar el tamaño de la muestra, para cada uno de los niveles, se deben utilizar las fórmulas 5, 11 y 12 consecutivamente.

Para el nivel 0 la celda crítica está definida por la intersección de la segunda línea y de la segunda columna de manera que el número de observaciones que respeta la condición de que la frecuencia esperada sea igual a cinco es:

$$n_0 = \frac{5}{0{,}30 \times 0{,}40} = 42$$

Generándose así, la tabla de frecuencias esperadas:

	M	M'	
R	17,3	7,7	25
R'	11,7	5,3	17
	29	13	42

Para el nivel 1,

$$n_0 = \frac{5}{0{,}20 \times 0{,}10 \times 0{,}30} = 833$$

Este valor se determina en función de la celda crítica del nivel 2. Esta celda se encuentra en la tabla de la izquierda y las frecuencias marginales mínimas que la determinan son 0,10 y 0,30. El producto de las marginales de la celda crítica, y de la probabilidad de distribución de esta tabla 0,20, entrega por resultado 0,006, en contraposición con el 0,032 que entrega el producto de las marginales de la celda crítica por la probabilidad de distribución en la otra tabla.

El resultado de la aplicación de este tamaño de muestra a las tablas de nivel 1, entrega los siguientes cuadros de frecuencias esperadas.

H

	M	M'	
R	5,0	45,0	50
R'	12,0	105,0	117
	17	150	167

H

	M	M'	
R	426,6	106,4	533
R'	106,4	426,6	133
	533	426,6	666

En el nivel 2 se tienen 4 tablas, y los productos de las marginales por las probabilidades de distribución del nivel 0 al 1 y del 1 al 2, se presentan en la tabla 7.

TABLA 7

Productos de celdas críticas por probabilidades de distribución:

0,0016	0,0048	0,0016	0,144

Las cifras de este cuadro están ordenadas según la disposición de las tablas originales. Así, el 0,0016 corresponde a la tabla que se ubica en el extremo izquierdo del segundo nivel, en tanto que el 0,144 a la tabla del extremo derecho en ese mismo nivel.

Aplicando la fórmula para determinar el tamaño de la muestra al nivel 2 y haciendo uso de la información contenida en la tabla anterior se tiene:

$$n_0 = \frac{5}{0{,}0016} = 3.125$$

Aplicando este tamaño de muestra para calcular las frecuencias esperadas se obtienen los cuadros:

HU

	M	M'	
R	5,0	45,0	50
R'	45,0	405,0	450
	50	450	500

HU

	M	M'	
R	34,8	15,2	50
R'	52,2	12,8	75
	87	38	125

H'U

	M	M'	
R	5,0	45,0	50
R'	20,0	180,0	200
	25	225	250

H'U'

	M	M'	
R	1.620	180	1.800
R'	405	45	450
	2.025	225	2.250

La estrategia planteada en la sección precedente se resume en la construcción de la tabla 7 y para calcular el tamaño de muestra que cumple con el criterio que las frecuencias esperadas sean por lo menos iguales a cinco, bastará con dividir 5 por el valor mínimo contenido en ella.

VI. Tamaño de muestra y análisis de asociación: Generalización *

La extensión de los resultados teóricos que hemos obtenido sobre un conjunto cualquiera de variables cualitativas pluricotómicas requiere de una simbología más compleja que la usada hasta el momento. Una simbología que permita un manejo fluido de las expresiones formales.

Simbolicemos las probabilidades de distribución por $P_{1,i+1;j}{}^{(i)}{}_{k}{}^{(i+1)}$ donde $i = 1,2,3,\ldots, m$; $j,k = 1,2,3,\ldots, q$ con la restricción que $k^{(i)} = j^{(i+1)}$.

De los cuatro subíndices que afectan a p los dos primeros, i e $(i + 1)$, nos señalan que las probabilidades de distribución conectan dos niveles consecutivos. Los subíndices $j^{(i)}$ y $k^{(i+1)}$ nos permiten ubicar la tabla *j* del nivel del origen *i* en relación a la tabla *k* del nivel de destino $(i + 1)$. La condición que $k^{(i)} = j^{(i+1)}$ nos indica que la tabla de destino en el nivel *i* es la de origen para el nivel $(i+1)$. De este modo, $P_{12;1}{}^{(1)}{}_{3}{}^{(2)}$ es la probabilidad de distribución que vincula a la tabla 3 del nivel 2 con la 1 del nivel 1.

La rama del árbol continúa a partir de la tabla 3 de destino con el nivel 2 que se convierte en la tabla de origen para el nivel 3. De acuerdo con la restricción $k^{(i)} = j^{(i+1)}$ al pasar del nivel 2 al 3 tenemos: $3^{(2)} = 3^{(3)}$ y las probabilidades de distribución que siguen en la misma rama serán: $P_{23;3}{}^{(3)}{}_{k}{}^{(4)}$, donde k asume un valor igual al número de categorías de la variable control incorporada al análisis.

$r^{i+1}_{min.,k}$ denotará la proporción marginal mínima en las líneas de la tabla *k* del nivel $(i + 1)$.

$r^{i+1}_{.min,k}$ simbolizará la proporción marginal mínima en las columnas de la tabla *k* del nivel $(i + 1)$.

En esta sección no se agrega ningún concepto en relación a los desarrollados previamente. Se ha incluido sólo con el propósito de mostrar que las ideas ya presentadas son fácilmente generalizables. El lector que tenga dificultades con el manejo de símbolos matemáticos puede pasarla por alto sin que ello tenga implicación conceptual alguna.

Sabemos, en virtud del procedimiento que hemos descrito en la sección anterior que el valor que interesa para determinar el tamaño de la muestra es el resultante del producto de las probabilidades de distribución multiplicadas por las probabilidades marginales que determinan la casilla crítica. Esto se puede expresar matemáticamente como:

$$13] \quad \prod_{i=1}^{m} P_{i,i+1;j}{}^{(i)}{}_{k}{}^{(i+1)} \qquad [r^{(i+1)}_{min.,k} \times r^{(i+1)}_{.min,k}]$$

en que $\prod_{i=1}^{m} P_{i,i+1;j}{}^{(i)}{}_{k}{}^{(i+1)}$ simboliza las probabilidades de distribución que se encuentran en la rama que lleva a la tabla k del nivel $(i+1)$.

Las operaciones indicadas por esta expresión pueden ser vertidas en una tabla equivalente a la número 7 que hemos presentado en la sección anterior, donde habrá tantos valores como tablas ubicadas en el último nivel.

En virtud de los planteamientos ya realizados sabemos que para determinar un tamaño de muestra que cumpla en todas las tablas con la condición que las frecuencias esperadas sean por lo menos iguales a cinco, se debe ubicar el producto mínimo de la rama por la casilla crítica, esto es equivalente a determinar:

$$\min \left\{ \pi \, P_{i,i+1;j}{}^{(i)}{}_{k}{}^{(i+1)} \qquad [r^{(i+1)}_{min.,k} \times r^{(i+1)}_{.min,k}] \right\}$$

en consecuencia el tamaño de muestra que se requerirá en la tabla original para que se respete el criterio que hemos impuesto resultará de:

$$14] \quad n_0 = \frac{5}{\min \left\{ \pi \, P_{i,i+1;j}{}^{(i)}{}_{k}{}^{(i+1)} \quad [r^{(i+1)}_{min.k} \times r^{(i+1)}_{.min,k}] \right\}}$$

Así llegamos a disponer de una fórmula general que nos permite establecer un tamaño de muestra que en todas las tablas cumple con el criterio que las frecuencias esperadas deben ser por lo menos iguales a cinco.

VII. Algunas consideraciones adicionales

Los procedimientos muestrales normalmente en uso para estimar parámetros descriptivos de una población demandan un conjunto mínimo de

información para proceder al cálculo del tamaño de la muestra. El investigador debe de disponer, por lo menos, de algunas estimaciones relativas a varianzas de las variables básicas de la investigación y conocer el tamaño de la población, estratos, conglomerados, unidades de primera o n-ésima etapa, según sea el diseño muestral que haya seleccionado como el más adecuado para abordar el problema que tiene entre manos. Esta información puede provenir, como ya hemos señalado, de otros muestreos que se hayan realizado, de datos censales o bien de muestras piloto o iluminación.

Por otra parte, las fórmulas tradicionales para determinar el tamaño de la muestra han sido construidas para estimar la media poblacional de *una variable*. En efecto, tanto el error máximo admisible como la varianza muestral (s^2) expresan relación con las características de una variable: s^2 simboliza la dispersión de una variable particular y si el muestreo tiene como objetivo investigar sobre un conjunto de variables tendremos también un conjunto de varianzas y al ser todas ellas distintas se generarán tantos tamaños de muestra como elementos tenga el conjunto; d que representa la discrepancia entre la media muestral y la poblacional, también se refiere a una diferencia para una *variable específica,* es evidente que el valor numérico que fije el investigador para d, dependerá de la naturaleza cualitativa o cuantitativa de la variable, por ejemplo, una diferencia de 0.4 puede ser muy grande si se trata de estimar tasas de desocupación y muy pequeña si la variable es el ingreso en sucres. Disponer de una serie de valores para s^2 y d es suficiente para justificar que en general habrá tantos tamaños de muestras como variables, sin embargo, podríamos considerar que no hay razón alguna para que todas las variables deban estimarse con los mismos niveles de confianza lo que resultaría en diferentes t para las distintas variables, hecho que refuerza la aseveración relativa a varios tamaños de muestra.

Para elegir entre los tamaños de muestra que resultan de la aplicación de las fórmulas es aconsejable, en primer lugar, reducir el universo de variables sólo a aquéllas que son consideradas básicas o centrales para la investigación. Una vez que se han eliminado las variables accesorias se puede seguir uno de los siguientes caminos: *a)* Seleccionar el mayor de entre todos los tamaños de muestras correspondientes a las variables básicas; *b)* Tomar como tamaño de muestra el promedio, el valor n mediano; el n modal. La primera alternativa presenta como ventaja principal que ese número de observaciones debería cubrir con exceso las exigencias técnicas (errores máximos admisibles y niveles de confianza) impuestas al proceso de estimación, pero puede entregar como resultado final un n excesivamente grande. Puede acontecer que el valor máximo de tamaño de muestra sea un valor extremo que no justifique un gasto adicional

significativo. El segundo camino tiene como inconveniente que el tamaño de muestra elegido relaja los requisitos técnicos para un conjunto de variables, aun cuando comparado con el criterio anterior es notoriamente menos exigente desde el punto de vista económico.

El procedimiento que hemos expuesto para determinar el número de observaciones que permita el cruce simultáneo de un conjunto de variables también implica, como hemos visto, la necesidad de seleccionar un tamaño de muestra entre los varios que resultan de multiplicar las proporciones de las casillas críticas de las tablas de último nivel, por las correspondientes probabilidades de distribución.

Además, el cálculo de n también demanda en este caso de un trabajo previo de estimación, pero en lugar de estimar las varianzas de las variables, se necesitan estimaciones de las probabilidades marginales y de distribución. La información necesaria para llevar a cabo las estimaciones que se necesitan para calcular n pueden obtenerse a través de censos, otras muestras o bien de la muestra de iluminación. Como se puede apreciar el problema de estimación previo al cálculo del tamaño de la muestra guarda un paralelismo estrecho entre el procedimiento propuesto y los usuales.

La diferencia básica entre uno y otro radica en las exigencias teóricas formuladas por la técnica. Cuando se trata de estimar medias y varianzas poblacionales, el papel de las proposiciones teóricas se reduce a delimitar los indicadores y variables sobre los cuales se debe recoger información. Por ejemplo, una investigación sobre formación del proletariado en el agro puede demandar información relativa a las relaciones de explotación, a las relaciones de propiedad y a las relaciones técnicas, de manera que se pone el acento en las proporciones * de arrendatarios y propietarios, proporciones de compradores y vendedores de la fuerza de trabajo y en el nivel de desarrollo de las fuerzas productivas.

La determinación del número de observaciones necesarias para realizar el análisis de asociación no sólo implica que las ideas teóricas sean capaces de delimitar las variables sobre las cuales se debería obtener información, sino que pone la exigencia al nivel del desarrollo de un plan de cruces previo a la selección de la muestra. Esto implica que la teoría debe ser lo suficientemente fuerte como para responder a tres órdenes de exigencias: *a)* El conjunto de variables a considerar, *b)* Las categorías probables que las compondrán y *c)* Cuántas y cuáles se cruzarán simultáneamente. Siguiendo con el ejemplo, podríamos pensar que el plan de cruces debería contener tablas formadas por el cruce simultáneo de las

* Por ejemplo, Lenin, V. I.: *El Desarrollo del Capitalismo en Rusia.* Moscú, Editorial Progreso, 1974, Cap. II.

tres variables de modo que las observaciones deberían tender a agruparse en tres categorías. Aquellas que pertenecen a la celda definida por: venta de fuerza de trabajo y no propiedad de medios de producción; compra de fuerza de trabajo, propietario de medios de producción, que usan alto nivel de desarrollo de las fuerzas productivas y ni compra ni vende fuerza de trabajo sino que usa la fuerza de trabajo familiar, propietario de los medios de producción y bajo nivel de desarrollo de las fuerzas productivas.

En la estrategia que proponemos sólo con el conocimiento del plan de cruces se sabrá qué probabilidades marginales y de distribución habrá que estimar y en consecuencia sólo a partir de él tendrá sentido preguntarse por el número de observaciones necesarias para llevar a cabo el análisis de asociación que demanda el esquema teórico.

VIII. A modo de conclusiones

La intención de este escrito ha sido la de poner en correspondencia la estrategia de análisis empírico que destila del pensamiento teórico, con el tipo de preguntas y consideraciones que surgen desde el ámbito de la teoría matemática del muestreo.

Hemos mostrado que las expresiones corrientemente utilizadas corresponden al criterio estadístico de garantizar que el error máximo admisible (discrepancia entre la media de la muestra y de la población) no supere cierto valor con un nivel de confianza prefijado. Es decir, las fórmulas que normalmente se usan para determinar tamaños de muestras se vinculan directamente a la estimación de promedios poblacionales. En consecuencia, lo más frecuente es que la solución al problema sea monótona con respecto a cualquier tipo de pregunta que se formule el investigador con respecto al número de observaciones. De este modo se rompe con la correspondencia que debiera existir entre los criterios provistos por la estadística y los que emergen del tipo de análisis que demanda una construcción teórica específica. Por un lado tenemos un cuerpo de conocimientos desarrollados en el interior de la teoría estadística cuya estructuración corresponde a una lógica interna y por otro, el manejo de un conjunto de categorías conceptuales que condicionan el tipo de análisis que más se adecúa y que debieran traducirse en ciertos criterios que además de delimitar el tipo de técnicas a usarse nos permitan derivar las exigencias de cantidad de información. Según esta óptica el origen del problema que hemos abordado se deposita en la no correspondencia entre los criterios meramente estadísticos que subyacen a la determinación del

tamaño de la muestra y aquellos que derivan de los requerimientos de información para realizar un análisis a través de tablas cruzadas.

Tal vez no esté de más señalar que aun dentro de la misma estadística matemática encontramos diversas fórmulas para determinar tamaños de muestras que responden a preguntas distintas de la simple estimación de medias poblacionales. Así por ejemplo,

$$n = \frac{\sigma^2(Z_\alpha + Z_\beta)}{(\mu_1 - \mu_0)^2}$$

es una fórmula que nos permite calcular un tamaño de muestra tal que se pueda someter a prueba una hipótesis nula en contra de una alternativa de una cola, de modo que los errores de tipo I y II alcancen valores prefijados.* En esta expresión σ^2 es la varianza de la poblacion. μ_1 y μ_0 son las medias poblacionales postuladas por las hipótesis alternativa y nula respectivamente. Z_α y Z_β son coeficientes que provienen de la curva normal y se calculan sobre la base del tamaño de error de tipo I (α) y el tamaño de error de tipo II (β).

La idea central que hemos expuesto es bastante simple: debe existir una correspondencia entre los criterios estadísticos que permiten el desarrollo de fómulas para determinar *n*, con aquellos que derivan del discurso teórico. Como hemos visto esta idea se encuentra avalada aun por los desarrollos dentro de la propia estadística matemática.

En este texto hemos examinado las consecuencias que se derivan de *un criterio particular* que tal vez pueda llegar algún día a gozar de alguna popularidad entre aquellos estudios que deban recurrir al análisis de asociación. Sin embargo, no debe pensarse que nuestra proposición constituye la única alternativa, todo dependerá del tipo de pregunta estadística que derive el investigador a partir de sus consideraciones de orden teórico. Así, por ejemplo, si deseamos realizar un estudio de contingencia con un conjunto de variables dicotómicas y además hemos optado por usar el coeficiente Q de Yule para medir la fuerza de la relación, podríamos pensar en establecer como criterio para calcular *n* el que la discrepancia entre el valor muestral y el poblacional no sea mayor que un valor previamente fijado con un determinado nivel de confianza.

La no correspondencia entre la naturaleza de la fórmula para calcular el tamaño de la muestra y las demandas técnicas que surgen del esquema teórico no es la única fuente desde la cual puede surgir una inadecuación entre el número de observaciones de que se dispone y las necesidades de información que derivan de la técnica de análisis estadístico a usarse.

* John E. Freund. *Mathematical Statistics*. Prentice Hall, 1962, p. 265.

Es común que en las investigaciones sociales se decida tomar una muestra tan grande como lo permitan las restricciones presupuestarias, las que normalmente son bastantes estrechas. No debe resultar extraño, entonces, que a veces se presentan a la consideración del lector análisis que contienen tan pocos datos que un cambio de una observación de una casilla a otra pueda hacer variar el sentido de la relación bajo estudio. Este hecho reviste especial gravedad en la medida que sabemos que las técnicas de categorización de variables difícilmente entregan normas precisas para establecer unívocamente los cortes. La sensibilidad de los hallazgos empíricos al escaso número de observaciones se puede obviar si el cruce de variables se lleva hasta el nivel en que se cumpla con los criterios exigidos por la estadística.

Este procedimiento partiría de los recursos económicos disponibles y en función de ellos limitaría los alcances del análisis, en circunstancias que pareciera que el interés del analista debiera consistir en el problema inverso: disponer de una idea respecto al número de observaciones que necesitaría para llevar a cabo la investigación que se ha propuesto. El riesgo de tomar un tamaño de muestra sobre la base del presupuesto puede llegar al extremo de que el número de observaciones no sea suficiente para ayudar en la construcción de las respuestas a las preguntas más elementales de la investigación.

En fin, el cálculo del tamaño de la muestra no se reduce simplemente a aplicar una fórmula cualquiera extraída desde un libro de texto, por reiterado que sea su uso, sino que hay que compatibilizar las restricciones presupuestarias con tamaños de muestras que permitan llevar a cabo los análisis empíricos que derivan de las consideraciones teóricas.