PROMECAFE

MANUAL PRACTICO PARA EL ANALISIS DE DATOS OBTENIDOS POR MUESTREO

Víctor Quiroga

IICA

ZONA NORTE

INSTITUTO INTERAMERICANO DE CIENCIAS AGRICOLAS – OEA

Esta publicación pertenece a una serie que tiene como propósito poner al alcance de los técnicos que trabajan en café, de caficultores progresistas y de agrónomos en general, artículos en castellano o traducciones de artículos en otros idiomas, que por su circulación restringida no alcanzan a llegar a un público más extenso y por su utilidad y valor merecen conocerse en los países que forman el PROMECAFE.

PUBLICACION MISCELANEA No. 214

ISSN 0534-5391

Programa Cooperativo para la Protección y Modernización de la Caficultura en

México, Centroamérica y Panamá

PROMECAFE

MANUAL PRACTICO PARA EL ANALISIS DE DATOS OBTENIDOS POR MUESTREO

Víctor Quiroga

CENTRO INTERAMERICANO DE DOCUMENTACION, INFORMACION Y COMUNICACION AGRICOLA – CIDIA
San José, Costa Rica
Junio 1979

IICA
PM–214

Quiroga, Víctor
Manuel práctico para el análisis de datos obtenidos por muestreo. San José, Costa Rica, CIDIA, 1979.
54 p. (IICA. Publicación Misceláneа, no. 214).

1. Análisis Estadístico. I. Título. II. Serie.

CDD 311.23
AGRIS U10

PRESENTACION

El muestreo de cafetales es una técnica cada vez más usada en nuestros países, tanto para detectar enfermedades y plagas como también para el pronóstico de cosechas.

Por la razón antes apuntada, el PROMECAFE realizó un seminario sobre ese tema al que asistieron funcionarios de los países del área y del Caribe.

Esa actividad se llevó a cabo con la colaboración del Instituto Mexicano del Café, en sus instalaciones situadas en Xalapa, Veracruz, México.

En dicho seminario se hizo evidente la necesidad de manuales que describan metodologías estadísticas aplicadas a la evaluación de muestras lo que nos induce a presentar este material técnico para el análisis e interpretación de resultados conseguidos por muestreo.

PROMECAFE
IICA-Sede Central
Coronado, San José, Costa Rica
Junio de 1979

CONTENIDO

Página

Página

LISTA DE CUADROS

1. INTRODUCCION

1.1 Conceptos Generales

Muestreo es la parte del Método Científico que trata de la recolección, tabulación y resumen de datos seleccionados de una población finita.

El investigador que se dedica a recolectar datos acerca de prácticas agrícolas, producción, comercialización, costo, etc., se enfrenta a la alternativa siguiente: efectuar una enumeración total de la población en estudio o tomar una muestra representativa, caso en que las mediciones se realizan en algunos individuos de la población. De tal enumeración total, el investigador tiene la esperanza de obtener su información completa y exacta; es decir, si él desea conocer el volumen de producción de frijol en determinada área geográfica, podría recolectar la información de la población dedicada a la producción de frijol por entrevista con cada uno de los agricultores, por razones económicas el investigador se conforma con una muestra representativa de agricultores de quienes obtiene la información deseada. Esto induce a establecer algunas diferencias y ventajas entre estas 2 formas de colección de datos

1.2 Diferencias entre Muestreo y Censo

Indicador	Muestreo	Censo
Costo	Muy bajo	Muy alto
Velocidad de recolección de datos	Alta	Lenta
Resumen y análisis	A corto plazo	A largo plazo
Precisión	Buena	Baja
Confiabilidad	Alta	Baja

1.3 Diseño de Muestreo y Diseño de Experimentos

Es también oportuno dejar establecidas las diferencias existentes entre estas 2 metodologías de colección de datos.

Indicador	Muestreo	Diseño de Experimentos
Propósito	Recolectar datos	Generar Datos
Población implicada	Finita	Infinita
Unidad informacional	Elegido al azar	Controlada por el invest.
Método estadístico	Estimación parámetros	Pruebas de hipótesis
Definición conceptual	Selección de una parte de población	Pregunta planeada

1.4 Definición de términos estadísticos

1.4.1 Población. Es el campo de estudio del fenómeno natural formado por toda la colección de unidades muestrales; es el conjunto universal de la variable que define el problema en estudio. Ejemplo, población de fincas, hogares, personas, etc. Se simboliza usualmente por letras latinas con subíndice.

$$Y_i = \{Y_1, Y_2, Y_3, \ldots\ldots, Y_i, \ldots\ldots\ldots, Y_N\}$$

Donde:

$$i = 1, 2, 3, \ldots\ldots\ldots, N$$

En general, N es usualmente grande y desde luego un número finito.

1.4.2. Muestra. Es una parte de la población, es un subconjunto del conjunto universal del fenómeno estudiado. Se simboliza en forma análoga a la definición anterior:

$$Y_i = \{Y_1, Y_2, Y_3, \ldots\ldots\ldots\ldots, Y_i, \ldots\ldots\ldots, Y_n\}$$

Es decir, utilizaremos en el Manual en notación siguiente:

Tamaño de la población .... N

Tamaño de la muestra .... n

Variable estudiada .... Y_i

1.4.3 Parámetros. Reciben este nombre, los valores típicos o descriptores de la población, se simbolizan por letras latinas mayúsculas; destacándose los siguientes:

a) Promedio poblacional ($\bar{Y}$).

$$Y = Y_1 + Y_2 + Y_3 + \ldots\ldots\ldots + Y_i + \ldots\ldots\ldots + Y_N$$

b) Promedio poblacional ($\bar{Y}$).

$$\bar{Y} = \frac{1}{N}(Y_1 + Y_2 + Y_3 + \ldots + Y_i + \ldots + Y_N) = \frac{1}{N}\Sigma Y_i$$

c) Media de medias de todas las muestras posibles de tamaño n.

$$\bar{\bar{Y}} = \frac{1}{\binom{N}{n}}(\bar{Y}_1 + \bar{Y}_2 + \ldots\ldots\ldots\ldots + \bar{Y}_{\binom{N}{n}})$$

d) Varianza poblacional de Y_i.

$$\sigma^2 = \frac{1}{N}\Sigma(Y - \bar{Y})^2$$

e) Varianza poblacional de las medias $\bar{Y}$.

$$\mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1}\right)$$

1.4.4 Definición de las variables. Con base en la población definida, se debe establecer el número de características que se considerarán en el muestreo; por ejemplo, para los 3 universos que siguen se puede obtener información con respecto a las características deseadas:

POBLACION	VARIABLES
Personas	Pesos, Estatura, etc.
Fincas	Area con maíz, ganado vacuno, etc.
Hogares	No. de Dependientes, ingreso por familia, etc.

1.4.5 Estimadores. Son los valores característicos que definen completamente la muestra, entre ellos tenemos:

a) Total de Y_i en la muestra

$$y = Y_1 + Y_2 + \ldots\ldots\ldots + Y_i + \ldots\ldots\ldots + Y_n$$

b) Promedio muestral.

$$\bar{y} = \frac{1}{n}(Y_1 + Y_2 + \ldots + Y_i + \ldots + Y_n) = \frac{1}{n}\Sigma Y_i$$

c) Varianza de la muestra Y_i

$$s^2 = \frac{1}{n-1}\Sigma(Y_i - \bar{Y})^2$$

d) La raíz cuadrada de la varianza recibe el nombre de desviación estándar. (s).

$$s = \sqrt{s^2}$$

e) Varianza muestral de $\bar{Y}$'s

$$s^2_{\bar{Y}} = \frac{s^2}{n} \cdot \left(\frac{N - n}{N - 1}\right)$$

f) La raíz cuadrada de la varianza de medias recibe el nombre de error estandar de la media o en forma más breve error estandar (E.E.)

$$EE = \sqrt{s^2_{\bar{Y}}} = s_{\bar{Y}}$$

1.4.6 <u>Estimado.</u> Es la información obtenida con base en el estimador muestral.

a) Estimado del total poblacional.

$$\hat{Y} = N\bar{y}$$

b) Estimado del promedio poblacional.

$$\hat{\bar{Y}} = \bar{y}$$

1.4.7 <u>Tipología de las estimaciones</u>

1) El <u>total</u> de la característica en la población; por ejemplo, el número de desocupados de un país, pesos de los universitarios, etc.

2) Proporción de la característica en la población, por ejemplo, Hectáreas con cultivos de maíz.

3) Razón entre 2 características de la población, por ejemplo, ingresos por familia y gastos en vestido.

1.4.8 Precisión de los estimadores. Es función del grado en que los estimadores aproximan a su valor esperado, se evaluan mediante: a) el coeficiente de variación comúnmente llamado error relativo y b) por el intervalo de confianza.

$$C.V. = \frac{s_{\bar{y}}}{\bar{y}} \cdot 100$$

$$L.C. = \bar{y} \pm 2s_{\bar{y}}$$

1.4.9 Errores ajenos al muestreo. En las investigaciones que implican metodologías de muestreo, casi siempre se cometen errores de medición o respuesta al momento de recolectar datos, los errores son más dramáticos cuando tienen sesgo sistemático como en el caso de encuestas con comerciantes que siempre subestimarán sus ventas por temor a leyes impositivas.

1.5 Tamaño de Muestra

De modo general cuanto más grande la muestra la precisión será mejor ya que ella es función del error estándar del estimador; sin embargo, se debe tener presente que a mayor tamaño de muestra aumenta la probabilidad de cometer errores ajenos al muestreo, por otra parte suben los costos del trabajo, habrá que transigir entre bajar costos y sacrificar precisión.

1.6 Tipos de Muestreo con base en Probabilidad

1.6.1 Muestreo con reposición. De la población con N unidades se selecciona los elementos muestrales con probabilidades iguales a 1/N, es.decir, se anota la selección y se devuelve a la población'

1.6.2 Muestreo sin reposición. De la población en estudio , se seleccionó el primer elemento con probabilidad 1/N, se anota la selección y se elimina de la población original, de este modeo el segundo y siguientes elementos son seleccionados con probabilidades 1/N, 1/N - 1, 1/N - 2 hasta 1/N - n + 1.

1.6.3 Muestreo con probabilidades de selección desiguales. Se recurre a una variable auxiliar (X), la misma que se supone correlacionada con la característica en estudio. En el Cuadro 1 se ilustra los tipos de muestreo con probabilidades iguales y desiguales.

Cuadro 1. Población de 5 fincas productoras de café

NUMERO	1970 - 1978		PROBABILIDAD $p_i = 1/N$	ESTIMACION $(y_i)/p_iN$	PROBABILIDAD $p_i = X_i/800$	ESTIMACION $N(Y_i)/p_iN$
	X_i	Y_i				
1	200	220	0.2	220	0.2500	176.000
2	300	350	0.2	350	0.3750	186.666
3	150	160	0.2	160	0.1875	170.666
4	100	110	0.2	110	0.1250	176.000
5	50	60	0.2	60	0.0625	192.00
	800	900		900		
PROMEDIO	160	180		180		180.26

1.7 Métodos de Estimación

1) Estimación por el total

$$\hat{y} = \bar{y} \cdot N$$

1.7.1 Estimación por diferencia.

$$\hat{y} = \bar{X} + K\ (Y - X)$$

1.7.2 Estimación por la razón.

$$\hat{y} = \frac{\hat{y}}{\hat{X}}\bar{X}$$

1.7.3 Estimación por regresión.

$$\hat{Y} = y - b(X - \bar{X})$$

Ejemplo 1. Métodos de estimación con base en una variable auxiliar

Cuadro 2. Tres métodos de estimación

NUMERO	X_i	Y_i	DIFERENCIA $\bar{X} + (Y-X)$	RAZON $y/X\ \bar{X}$	REGRESION $Y-b(X-\bar{X})$
1	200	220	180	176.000	173.52
2	300	350	210	186.666	187.29
3	150	160	170	170.666	171.62
4	100	110	170	176.000	179.73
5	50	60	170	192.000	187.84
PROMEDIO	160	180	180	180.26	180.00

Con los datos del Cuadro siguiente se estima b.

Cuadro 3. Sumas para calcular el coeficiente de regresión lineal b

	X	Y	X^2	XY	Y^2
	200	220	40000	44000	48400
	300	350	90000	105000	122500
	150	160	22500	24000	25600
	100	110	10000	11000	12100
	50	60	2500	3000	3600
SUMAS	800	900	165000	187000	212200

$$b = \frac{187000 - (800)(900)/5}{165000 - 800^2/5} = \frac{43000}{37000} = 1.16216216$$

1.8 Ilustración

Calcular los parámetros de una población finita imaginaria de tamaño $N = 4$, Y_i = producción en Toneladas por Hectárea de producto.

$$Y_i = \{Y_1, \ldots, Y_N\} = \{Y_1, Y_2, Y_3, Y_4\} = \{0, 2, 1, 3\}$$

Total en la población.

$$Y = Y_1 + Y_2 + Y_3 + Y_4 = 0 + 2 + 1 + 3 = 6 \text{ Ton.}$$

Media poblacional

$$\bar{Y} = \frac{1}{N} \Sigma Y_i = \frac{0 + 2 + 1 + 3}{4} = 1.5 \text{ Ton/Ha.}$$

Media poblacional de todas las muestras posibles de tamaño $n = 2$.

El número de muestras posibles de tamaño 2 de una población con 4 individuos, se calcula utilizando la definición dada por el análisis combinatorio.

$$\binom{N}{n} = \frac{N!}{n!\,(N-n)!} = \frac{4!}{2!\,2!} = 6$$

Las muestras posibles y sus promedios se tabulan a continuación.

Cuadro 4. Muestras posibles de una población N = 6

No.	Muestras	Medias
1	(0,2)	1.0
2	(0,1)	0.5
3	(0,3)	1.5
4	(2,1)	1.5
5	(2,3)	2.5
6	(1,3)	2.0

Varianza de la población.

$$\sigma^2 = \frac{(0-1.5)^2 + (2-1.5)^2 + (1-1.5)^2 + (3-1.5)^2}{4} = \frac{5}{4} = 1.25$$

Varianza de las medias posibles.

$$\sigma^2_{\bar{Y}} = \frac{(1.0-1.5)^2 + (0.5-1.5)^2 + (1.5-1.5)^2 + (1.5-1.5)^2 + (2.5-1.5)^2 + \ldots}{6}$$

$$= 0.416666$$

La varianza de las medias se puede calcular en función de la varianza del modo siguiente:

$$\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n} \cdot \left(\frac{N-n}{N-1}\right) = \frac{1.25}{2} \cdot \left(\frac{4-2}{4-1}\right) = \frac{2.5}{6} = 0.416666$$

Ejemplo 2. **Calcular los estimadores más notables de la muestra.**

$$Y_i = \{3,\ 1,\ 4,\ 3,\ 8,\ 3,\ 4,\ 0\}$$

Media aritmética.

$$\bar{y} = \frac{1}{n}\,\Sigma Y_i = \frac{1}{8}\,(26) = 3.25$$

Varianza de la muestra.

$$s^2 = \frac{1}{n-1}\,\{\Sigma Y_i^2 - (\Sigma Y)^2/n) = \frac{1}{7}\,(124 - (26)^2/8) = 5.642857$$

Varianza de las medias.

$$s_{\bar{y}}^2 = \frac{s^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{5.642857}{2}\left(\frac{4-2}{4-1}\right) = 1.8809$$

Error estándar de la media.

$$s_{\bar{y}} = \sqrt{1.8809} = 1.3714$$

2. MUESTREO IRRESTRICTO AL AZAR. (Variables cuantitativas)

2.1 Usos. Se recomienda el uso de este diseño en los casos de relativa homogeneidad de las unidades muestrales. La característica de la variable que se mide debe ser función unicamente de la variabilidad natural; es decir, no existe otra fuente de variación además de la atribuible al error de muestreo.

2.2 Marco. Es la lista organizada de todos los componentes de la población; es la presentación organizada del conjunto universal definido por la variable Y_i.

$$Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots\ldots Y_N\} = \text{marco}$$

2.3 Muestra. Es la colección de n unidades muestrales obtenidas del marco definido anteriormente.

$$Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots\ldots Y_n\} = \text{muestra}$$

2.4 Selección de la muestra. Para que la muestra sea representativa de la población, es aconsejable seguir los pasos siguientes:

2.41 Levantar la lista de todas las unidades muestrales, bien podría ser marco de área o marco de lista, según la situación.

2.42 Numerar en forma conveniente las unidades muestrales

2.43 Obtener las n unidades muestrales usando la tabla de números al azar.

2.44 Recolectar en el campo los datos para todas las variables definidas, según el cuestionario elaborado para la investigación del caso.

2.5 Análisis de la Información. Se realiza la evaluación numérica de los estimadores de acuerdo a las definiciones siguientes:

$$\bar{y} = \frac{1}{n} \Sigma Y_i$$

$$s^2 = \frac{1}{n-1}\,\Sigma(Y_i - \bar{y})^2$$

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

$$s = \sqrt{s^2}$$

$$s_{\bar{y}} = \sqrt{s_{\bar{y}}^2}$$

2.6 Estimados.

$$\hat{Y} = \bar{y}\,.\,N$$

$$\hat{\bar{Y}} = \bar{y}$$

$$s_{\hat{Y}}^2 = s_{\bar{y}}^2\, N^2$$

$$s_{\hat{Y}} = \sqrt{s_{\hat{Y}}^2}\; N$$

2.7 Parámetros.

$$Y = \Sigma Y_i$$

$$\bar{Y} = \frac{1}{N}\,\Sigma Y_i$$

$$\sigma^2 = \frac{1}{N}\,\Sigma(Y_i - \bar{Y})^2$$

El grado de confianza que puede depositarse en los estimados, depende fundamentalmente de la variación en las observaciones Y_i. Así, si los Y_i son muy parecidos en magnitud a causa de su escasa variación natural, una muestra pequeña arroja bastante luz acerca de la media poblacional. Por el contrario, si la variabilidad Y_i es grande, una muestra pequeña proveerá un estimado poco confiable.

Los grados de libertad (n-1) definidos al evaluar la varianza, constituye un indicador del grado de confiabilidad depositado en el estimado, ya que una varianza calculada con base en 1 grado de libertad (tamaño de muestra n=2) es poco confiable, comparado con otro que implica 20, 30, 60 ó más grados de libertad.

2.8 Límites de confianza del promedio. La distribución contínua de 't' se define con el cociente siguiente:

$$\pm t = \frac{\bar{y} - \bar{Y}}{s_{\bar{y}}}$$

Despejar $\bar{Y}$ por trasposición de términos en la ecuación

$$\bar{Y} = \bar{y} \pm (s_{\bar{y}})(t)$$

La doble desigualdad del lado derecho obliga a usar la expresión de desigualdad. siguiente:

$$\bar{y} - (s_{\bar{y}})(t) \leqslant \bar{Y} \leqslant \bar{y} + (s_{\bar{y}})(t)$$

Estos límites sin embargo, no son absolutos, porque aún existe la posibilidad de una muestra que arroje un estimado fuera del rango definido anteriormente, por lo que estos límites se los expresa en términos probabilísticos.

$$\text{Prob}\ \{\ \bar{y} - (s_{\bar{y}})(t_{\alpha}) \leqslant \bar{Y} \leqslant \bar{y} + (s_{\bar{y}})(t_{\alpha})\ \} = 1 - \alpha$$

2.9 Límites de confianza del total. Se evalua con base en la definición de la variable estandar t.

$$\pm t = \frac{\hat{Y} - Y}{s_{\hat{Y}}}$$

Despejando Y en el miembro derecho de la igualdad tenemos:

$$Y = \hat{Y} \pm (s_{\hat{Y}})(t)$$

$$\hat{Y} - (s_{\hat{Y}})(t) \leq Y \leq \hat{Y} + (s_{\hat{Y}})(t)$$

$$\text{prob } \{\hat{Y}-(s_{\hat{Y}})(t_{\alpha}) \leq Y \leq \hat{Y} + (s_{\hat{Y}})(t_{\alpha})\} = 1 - \alpha$$

Para evaluar los límites de confianza, se elige t para un determinado nivel α (0.10, 0.05, 0.01, etc).

Ejemplo 2.1 Suponer que se va a evaluar un bosque de Pinus sp. que abarca imaginariamente 2000 Hectáreas de terreno. El propósito de la investigación es estimar el volumen promedio de madera por Ha y volumen total en el bosque.

La unidad muestral es de una hectárea sobre la que se evalua el volumen maderable.

Metodología.

i) Dividir el bosque en 2000 unidades muestrales y enumerarlos consecutivamente, seleccionar una muestra con n=20 unidades, haciendo uso de la tabla de números al azar.

ii) Recolectar datos en cada una de las unidades y presentar en forma tabular.

Cuadro 5. Volumen de madera en m^3/unidad muestral

UNIDAD	VOLUMEN	UNIDAD	VOLUMEN
1	8.0	11	9.0
2	7.0	12	8.0
3	6.0	13	7.0
4	9.0	14	6.0
5	10.0	15	9.0
6	7.0	16	10.0
7	8.0	17	6.0
8	9.0	18	7.0
9	6.0	19	8.0
10	8.0	20	8.0

iii) Cálculo de estimadores:

$$\bar{y} = \frac{8.0 + 7.0 + \ldots\ldots\ldots\ldots\ldots + 8.0}{20} = 7.8/\text{Ha.}$$

$$s^2 = \frac{(8.0-7.8)^2 + \ldots\ldots + (8.0-7.8)^2}{19} = 1.6421$$

$$s^2_{\bar{y}} = \frac{1.6421}{20} \cdot \frac{2000 - 20}{2000 - 1} = 0.0813$$

$$s_{\bar{y}} = 0.285$$

iv) Cálculo de estimados:

$$\hat{\bar{Y}} = 7.8/\text{Ha.}$$

$$\hat{Y} = (7.8)(2000) = 15600.0$$

$$s_{\hat{Y}} = (0.285)(2000) = 570$$

v) Límites de confianza para los parámetros Y, $\bar{Y}$.

prob {15600-(570)(2.903) $\leq$ Y $\leq$ 15600 + (570)(2.093)} = 1 - 0.05

prob (14406 $\leq$ Y $\leq$ 16793) = 0.95

prob {7.8 - (0.285)(2.093) $\leq \bar{Y} \leq$ 7.8 + (0.285)(2.093)} = 1 - 0.05

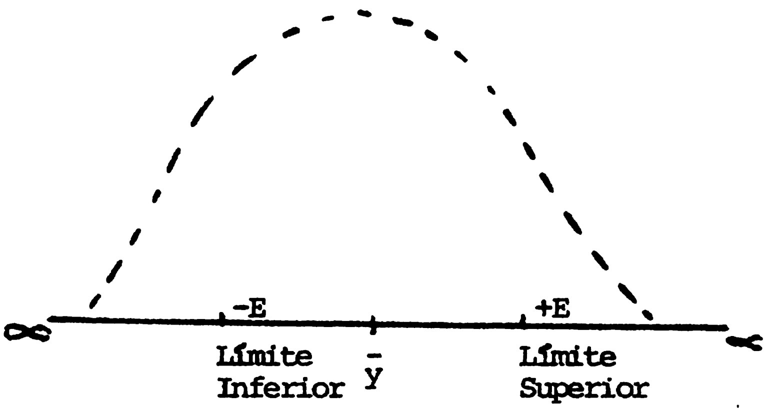
prob (7.20 $\leq \bar{Y} \leq$ 8.40) = 0.95

2.10 Tamaño de Muestra. Al planear un diseño de muestreo irrestricto al azar, se llega al punto crítico de tomar alguna decisión con respecto al tamaño de la muestra, una muestra excesivamente grande, implica un considerable derroche de recursos, mientras que una muestra muy pequeña minimiza la utilidad y bondad de los resultados. La decisión no siempre es satisfactoria, ni definitiva, ni única, por lo que se deberá transigir entre costos mínimos y máxima confiabilidad; es decir, a mayor número de observaciones mayor costo y mejor precisión, de otro modo, a menor número de observaciones menor costo y poca confiabilidad.

La definición de límites de confianza, establece:

$$\bar{y} - (s_{\bar{y}})(t) \leq \bar{Y} \leq \bar{y} + (s_{\bar{y}})(t)$$

cuya representación esquemática, conduce a definir el semi intervalo ±E.



Es decir:

$$E = (s_{\bar{y}})(t)$$

$$E^2 = (s^2_{\bar{y}})(t^2) \qquad [1]$$

Donde:

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

Además por ser N un valor grande el factor de corrección por población finita puede transformarse así:

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

Reemplazar este resultado en |1| para despejar n.

$$E^2 = \frac{s^2}{n}\left(1-\frac{n}{N}\right) t^2$$

$$\frac{E^2}{t^2} = \frac{s^2}{n}\left(1-\frac{n}{N}\right)$$

$$n = \frac{1}{\frac{E^2}{t^2 s^2} + \frac{1}{N}}$$

Si N es suficientemente grande.

$$n = \frac{t^2 s^2}{E^2} \qquad |2|$$

Esta fórmula que se utiliza para calcular el tamaño de muestra, depende de una propiedad peculiar del fenómeno en estudio (s^2), y de un valor cuasi subjetivo (E) que es la magnitud del error que se admite en la estimación correcta de $\bar{Y}$.

Desde que aún existe la posibilidad de la ocurrencia de una muestra desafortunada cuyo error (E) sea mayor que el esperado, se evalúa la expresión (n) en términos probabilísticos y para cualquier valor del nivel α ; inicialmente con infinito número de grados de libertad y luego por proceso de iteración hasta estabilizar n.

Ejemplo 2.2 De cierta población de tamaño n = 8000 cuya variable Y_i se desea medir, se conoce la varianza $s^2 = 81.0$. Cuál es el tamaño de muestra que con seguridad del 95% nos permita conocer el estimado con un error admisible de solamente 3 unidades? Es decir, si la media se espera sea 50.0 unidades, ésta no debe exceder los límites de: 50.0 ± 3.

1° iteración t con $\alpha = 0.05$; g.l. = ∞

$$n = \frac{1}{\frac{3^2}{(1.96)^2 (9.0)^2} + \frac{1}{8000}} = 34.4$$

2°. iteración t con $\alpha = 0.05$; g.l. = 30

$$n = \frac{1}{\dfrac{3^2}{(2.042)^2(9.0)^2} + \dfrac{1}{8000}} = 37.3$$

Es frecuente el uso de la fórmula alternativa siguiente para el cálculo del tamaño de muestra n.

$$n = \frac{t^2(CV)^2}{e^2}$$

Donde:

CV = Coeficiente de variación

e = error admisible porcentual

Con base en los datos del ejemplo 2.2 se obtiene el siguiente resultado

$$n = \frac{(1.96)^2(-18)^2}{0.06} = 34.5$$

3. MUESTREO PARA ESTIMAR PROPORCIONES Y PORCENTAJES (Var. cualitativas)

3.1 Generalidades. Este diseño de muestreo, es una extensión de Muestreo Irrestricto al Azar aplicado a variables cualitativas y cuantitativas discretas. En general las proporciones se presentan como frecuencia del total referido a la unidad, por lo que resulta de mayor interés conocer la proporción relativa antes que el total o su promedio. Ejemplo, proporción de desempleados en la población, el ingreso familiar a cierto nivel, proporción de personas en favor de cierto cambio, proporción de casas comerciales interesadas en introducir un nuevo producto, etc.

Suponer que las unidades muestrales en una población cualquiera, se dividen en 2 clases mutuamente excluyentes:

- La clase A que contiene todas las unidades que poseen el atributo bajo estudio.
- La clase B formado por los elementos que no poseen el atributo.

Sea la variable Y_i proveniente de una población finita de tamaño N.

$$Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots\ldots\ldots, Y_i, \ldots\ldots\ldots\ldots Y_N \}$$

Con categorías: $A : \text{si } Y_i \subset A$

$B : \text{si } Y_y \subset B$

es decir:

$$Y = \{ A \cup B \}$$

De esta población, se extrae una muestra de tamaño n.

$$Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots\ldots\ldots, Y_i, \ldots\ldots\ldots\ldots Y_n \}$$

Con categoría: a : si $y_i \subset A$

b : si $y_i \subset B$

La probabilidad de que en una muestra cualquiera de tamaño n, seleccionada al azar de una población N ocurra exactamente: a de la categoría A y b de la clase B, se define por:

$$\text{prob (a)} = \frac{\binom{A}{a}\binom{B}{b}}{\binom{N}{n}}$$

que es precisamente la distribución hipergeométrica, donde a toma valores { 0, 1, 2, ..., etc. }.

Ejemplo 3.1 Un frasco contiene 3 bolitas de tipo A y 4 bolitas de tipo B. Extraer una muestra de tamaño n = 2 con el propósito de evaluar: a) probabilidad de que ninguna sea A. b) La muestra tiene una A. c) La muestra está formada de dos bolitas tipo A.

$A = 3$

$B = 4$

$N = A + B = 7$

$n = 2$

$a = \{ 0, 1, 2 \}$

$$p(a = 0) = \frac{\binom{3}{0}\binom{4}{2}}{\binom{7}{2}} = \frac{6}{21} \quad \text{(probabilidad de que ninguna sea A)}$$

$$p(a = 1) = \frac{\binom{3}{1}\binom{4}{1}}{\binom{7}{2}} = \frac{12}{21} \quad \text{(probabilidad de que una de ellas sea A)}$$

$$p(a = 2) = \frac{\binom{3}{2}\binom{4}{0}}{\binom{7}{2}} = \frac{3}{21} \quad \text{(probabilidad de que ambas son A)}$$

$$p(a = 0) + p(a = 1) + p(a = 2) = 1.0$$

3.2 Evaluación de Estimadores. Se realiza utilizando una función indicadora que asigne el código 1 si Y_i pertenece a A, asigne un 0 si Y_i pertenece a B; es decir:

$$Y_i = 1 \quad \text{si } Y_i \subset A$$

$$Y_i = 0 \quad \text{si } Y_i \subset B$$

$$Y_i = \{ 1, 1, 1, 1, 1, \ldots\ldots\ldots\ldots, 0, 0, 0, 0 \}$$

De este arreglo, es interesante solamente la primera parte formada por el arreglo de los números uno; la segunda, se elimina porque no participará en los cálculos (no se toma en cuenta ningún cero).

3.21 Estimación de la proporción.

$$\bar{y} = \frac{1}{n} \Sigma Y_i$$

$$\Sigma Y_i = 1 + 1 + 1 + 1 + \ldots\ldots\ldots\ldots\ldots = a$$

$$\bar{y} = \frac{a}{n}; \text{ pero } \bar{y} = p$$

donde: $p = \frac{a}{n}$; como $p + q = 1$

$$q = 1 - p$$

$$q = 1 - \frac{a}{n}$$

$$a = p \, . \, n$$

3.2.2 Estimación de la varianza.

$$s^2 = \frac{1}{n-1}\left(\Sigma Y_i^2 - (\Sigma Y_i)^2/n\right)$$

Puesto que Y_i toma unicamente los valores 1 se cumple:

$$\Sigma Y_i = 1 + 1 + 1 + \ldots\ldots\ldots\ldots = a$$

$$\Sigma Y_i^2 = 1^2 + 1^2 + 1^2 + \ldots\ldots\ldots\ldots = a$$

Luego la varianza de la proporción se define así:

$$s^2 = \frac{1}{n-1}\left(a - \frac{a^2}{n}\right)$$

$$= \frac{a}{n-1} - \frac{a^2}{(n-1)(n)}$$

$$= \frac{a(n-a)}{n(n-1)}$$

$$= \frac{p(n-a)}{n-1}$$

$$= \frac{pn\left(1 - \frac{a}{n}\right)}{n-1}$$

$$s^2 = \frac{npq}{n-1}$$

La varianza de todas las proporciones posibles se estima mediante:

$$s_p^2 = \frac{s^2}{n}\left(\frac{N-n}{N-1}\right)$$

Reemplazando s^2 por la definición anterior:

$$= \frac{\frac{npq}{n-1}}{n}\left(\frac{N-n}{N-1}\right)$$

$$= \frac{pq}{n-1}\left(\frac{N-n}{N-1}\right)$$

La varianza de la proporción de A en la población tiene por estimador

$$s^2_{\hat{A}} = N^2\, s^2_{p}$$

3.3 Los parámetros estudiados son:

$$A = NP$$

$$P = \frac{A}{N}$$

3.4 Los estimados son:

$$\hat{A} = Np$$

$$\hat{p} = \frac{a}{n}$$

3.5 Límites de confianza.

$$\text{prob } \{\hat{A} - (s_{\hat{A}})(t_\alpha) \leq A \leq \hat{A} + (s_{\hat{A}})(t_\alpha)\} = 1 - \alpha$$

$$\text{prob } \{\hat{P} - (s_p)(t_\alpha) \leq P \leq p + (s_{\bar{p}})(t_\alpha)\} = 1 - \alpha$$

Ejemplo 3.2 De una población finita N = 960 manzanas, se toma una muestra n = 100, de éstas, 17 están atacadas por gusanos. Se desea evaluar los límites de confianza al 95% de la verdadera proporción de manzanas atacadas.

$N = 960$

$n = 100$

$a = 17$

$p = 0.17$

$q = 0.83$

$$s_p^2 = \frac{p\,q}{n-1}\left(\frac{N-n}{N-1}\right)$$

$$= \frac{(0.17)\,(0.83)}{99}\left(\frac{100-17}{100-1}\right)$$

$$= 0.001195$$

$$s_p = 0.034567$$

$$\text{prob} = \{(0.17)-(0.0357)(1.96) < P < 0.17 + (0.0357)(1.96)\} = 95\%$$

$$\text{prob} = (0.10 \leq P \leq .24) = 95\%$$

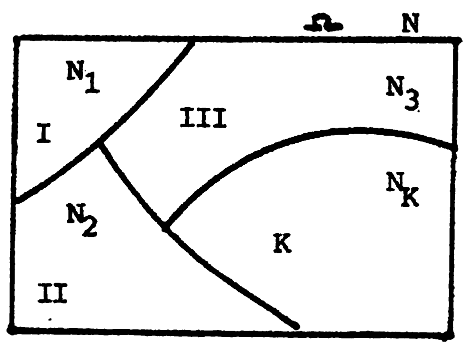
Tamaño de muestra. En forma similar al muestreo irrestricto al azar |2| se elabora la relación para calcular el tamaño de muestra.

$$n = \frac{t^2pq}{E^2} \qquad [3]$$

4. ESTRATIFICADO

4.1 Por qué estratificar? Una lista organizada, a veces no es suficiente para construir un diseño de muestreo si se desea aumentar la precisión de los estimadores, sobre todo cuando Y_i tiene un rango de variación muy grande.

4.2 Metodología del estratificado. La población se divide en grupos bien definidos, que reciben el nombre de estratos; graficamente se representa como el conjunto universal siguiente:



En general, se pueden indicar 3 tipos de estratificación.

- Grupos naturales: clima, suelo, topografía, etc.
- Divisiones político administrativas: provincia, cantón, municipio.
- Grupos artificiales: clases por tamaño de finca, edad, etc.

4.3 Notación. La población definida por la variable Y_i se representa así:

$$Y_i = (Y_1, Y_2, \dots\dots\dots, Y_i, \dots\dots\dots Y_N)$$

La misma después de la estratificación toma las expresiones:

$$\text{Estrato I } Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots, Y_i, \ldots\ldots\ldots\ldots, Y_{N_1} \}$$
$$\text{Estrato II } Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots, Y_i, \ldots\ldots\ldots\ldots, Y_{N_2} \}$$
$$\vdots \qquad \vdots$$
$$\text{Estrato K } Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots, Y_i, \ldots\ldots\ldots\ldots, Y_{N_k} \}$$

En forma similar la notación para muestra es:

$$\text{Estrato I } Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots, Y_i, \ldots\ldots\ldots\ldots, Y_{n_1} \}$$
$$\text{Estrato II } Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots, Y_i, \ldots\ldots\ldots\ldots, Y_{n_2} \}$$
$$\vdots \qquad \vdots$$
$$\text{Estrato K } Y_i = \{ Y_1, Y_2, \ldots\ldots\ldots, Y_i, \ldots\ldots\ldots\ldots, Y_{n_k} \}$$

4.4 Parámetros. El promedio estratificado es una información combinada o ponderada por la magnitud e importancia de cada estrato.

$$\bar{Y} \text{ compuesta de:} \begin{cases} \bar{Y}_1 \text{ del estrato } 1 \\ \bar{Y}_2 \text{ del estrato } 2 \\ \vdots \qquad\qquad \vdots \\ \bar{Y}_k \text{ del estrato } k \end{cases}$$

El total de los k estratos también es función de la importancia o peso relativo de cada estrato.

$$Y \text{ compuesta de:} \begin{cases} Y_1 \text{ del estrato } 1 \\ Y_2 \text{ del estrato } 2 \\ \vdots \qquad\qquad \vdots \\ Y_k \text{ del estrato } k \end{cases}$$

4.5 Estimadores. En forma similar los estimadores se calculan teniendo en cuenta los pesos de los estratos.

$$\bar{y}_{estrato} \text{ compuesta de: } \begin{cases} \bar{y}_1 \text{ del estrato 1} \\ \bar{y}_2 \text{ del estrato 2} \\ \vdots \qquad\qquad \vdots \\ \bar{y}_k \text{ del estrato k} \end{cases}$$

$$s_{\bar{y}estrato} \text{ compuesta de: } \begin{cases} s_{\bar{Y}} \text{ del estrato 1} \\ s_{\bar{Y}} \text{ del estrato 2} \\ \vdots \qquad\qquad \vdots \\ s_{\bar{Y}} \text{ del estrato k} \end{cases}$$

Separadamente en cada estrato, se utilizan las fórmulas definidas en el muestreo irrestricto al azar.

$$\bar{y}_j = \frac{1}{n_j} \Sigma Y_i \qquad\qquad j = \{ 1, 2, \ldots\ldots\ldots, k \} = \text{Estratos}$$

$$s_j^2 = \frac{1}{n_j - 1} \Sigma (Y_i - \bar{y}_j)^2$$

$$s_{\bar{Y}_j}^2 = \frac{s_j^2}{n} \left(\frac{N_j - n_j}{N_j - 1} \right)$$

Esto implica que todos los cálculos aritméticos deben efectuarse independientemente para cada estrato. Luego se combina la anterior información para obtener los estimadores del estratificado de la manera siguiente:

$$\bar{Y}_{estrato} = \Sigma \frac{N_j}{N} \bar{Y}_j$$

$$s_{\bar{y}estrato} = \Sigma \frac{N_j}{N} s_{\bar{Y}_j}$$

4.6 Estimados:

$$\hat{Y} = \Sigma \bar{y}_j N_j$$

$$s_{\hat{Y}} = \Sigma s_{\bar{y}} N_j$$

4.7 Límites de confianza.

$$\text{prob } \{\hat{Y} - (s_{\hat{y}estrato})(t) \leq Y \leq \hat{Y} + (s_{\hat{Y}estrato})(t)\} = 1 - \alpha$$

$$\text{prob } \{\bar{y} - (s_{\bar{y}estrato})(t) \leq \bar{Y} \leq \bar{y} + (s_{\bar{y}estrato})(t)\} = 1 - \alpha$$

Ejemplo 4.1 Suponer una población hipotética formada por N = 200 unidades observacionales y dividida en 3 estratos, de las que se obtienen 3 muestras de tamaño $n_j = \{5, 4, 6\}$. Se desea estimar los parámetros poblacionales con seguridad de 90%.

Cuadro 6. Datos tabulados para evaluar estimadores

Estrato	N_j	n_j	Producto	N_j/N	$\bar{y}_j$
I	40	5	{16, 15, 14, 15, 17}	0.20	15.40
II	50	4	{13, 14, 16, 15 }	0.25	14.50
III	110	6	{12, 10, 9, 12, 10, 9}	0.55	10.33
	N = 200	15 = n		Σ = 1.00	

Continuación de Cuadro 6.

s_j	$s_{\bar{y}_j}$	$\bar{y}_j(N_j/N)$	$s_{\bar{y}_j}(N_j/N)$	$\bar{y}_j.N_j$	$s_{\bar{y}_j}.N_j$
1.140175	0.457604	3.0800	0.091521	616	18.30416
1.290994	0.605977	3.6250	0.151494	725	30.29880
1.36326	0.532187	5.6815	0.292703	1136	58.54057
Σ:		12.3865	0.535718	2477	107.1435

Luego:

$\bar{Y}_{estrato} = 12.3865$ $\hat{Y}_{estrato} = 2477$

$s_{\bar{y}estrato} = 0.535718$ $s_{\hat{Y}estrato} = 107.1435$

Límites de confianza.

prob $\{\bar{y} - (s_{\bar{y}})(t) \leq \bar{Y} \leq \bar{y} + (s_{\bar{y}})(t)\} = 1 - \alpha$

prob $\{12.38 - (0.536)(1.782) \leq \bar{Y} \leq 12.38 + (0.536)(1.782)\} = 90\%$

prob $(11.42 \leq \bar{Y} \leq 13.34) = 90\%$

prob $\{\hat{Y} - (s_{\hat{Y}})(t) \leq Y \leq \hat{Y} + (s_{\hat{Y}})(t)\} = 1 - \alpha$

prob $\{2477 - (107)(1.78) \leq Y \leq 24.77 + (107)(1.78)\} = 1 - \alpha$

prob $(2286 \leq Y \leq 2667)$

4.8 Tamaño de muestra dentro de cada estrato. De modo general, se presenta estas 2 situaciones.

- No se tiene ninguna información previa
- Se conoce alguna información: varianza, costo por estrato, etc.

4.81 El tamaño de muestra n_j es proporcional a la magnitud del estrato Recibe también el nombre de afijación proporcional e implica que la intensidad muestral es proporcional al tamaño del estrato N_j; es decir, si cierto estrato está formado por el 50% de las observaciones, la muestra n_j también debe estar formada por el 50% del total n. El cálculo se realiza usando la expresión:

$$n_j = n \frac{(N_j/N)}{\Sigma(N_j/N)}$$

El denominador suma 1.0 por lo que se tiene la fórmula simplificada

$$n_j = n(N_j/N)$$

Ejemplo 4.2 Suponer cierta población, estratificada en k estratos, para calcular el tamaño de muestra en cada estrato.

Cuadro 7. Cálculo del tamaño de muestra por estrato.

Estrato	N_j	Fracción N_j/N	Tamaño n_j
I	N_1	N_1/N	$n_1 = n(N_1/N)$
II	N_2	N_2/N	$n_2 = n(N_2/N)$
⋮			
k	N_k	N_k/N	$n_k = n(N_k/N)$
	$\Sigma N_j = N$		$\Sigma n_j = n$

4.8.2 El tamaño de la muestra es proporcional a la varianza. Recibe también el nombre de afijación de Neymann, el tamaño de muestra en cada estrato es proporiconal a la varianza y al tamaño del estrato, se calcula así:

$$n_j = n \cdot \frac{(N_j/N)\ (s_j^2)}{\Sigma (N_j/N)\ (s_j^2)}$$

Ejemplo 4.3 Afijación de Neymann.

Cuadro 8. Cálculo del tamaño de muestra por Neymann.

Estrato	N_j	s_j^2	$(N_j/N)\ s_j^2 = f_j$	n_j
I	N_1	s_1^2	$(N_1/N)\ s_1^2 = f_1$	$n_1 = nf_1$
II	N_2	s_2^2	$(N_2/N)\ s_2^2 = f_2$	$n_2 = nf_2$
⋮	⋮	⋮	⋮	⋮
k	N_k	$s^2{}_k$	$(N_k/N)\ s_k^2 = f_k$	$n_k = nf_k$
	$\Sigma N_j = N$			$\Sigma n_j = n$

4.8.3 El tamaño de la muestra es proporcional al costo. Se conoce con el nombre de afijación óptima, es directamente proporcional al tamaño del estrato y a la varianza de la misma e inversamente proporcional al costo, así:

$$n_j = n \frac{(N_j/N)\,(s_j/C_j)}{\Sigma(N_j/N)\,(s_j/C_j)}$$

Ejemplo 4.4 Afijación óptima.

Cuadro 9. Cálculo del tamaño de muestra para afijación óptima.

Estrato	N_j	$(s_j/C_j).(N_j/N)=f_j$	n_j
I	N_1	$(s_1/C_1).(N_1/N)=f_1$	$n_1 = nf_1$
II	N_2	$(s_2/C_2).(N_2/N)=f_2$	$n_2 = nf_2$
⋮			
k	N_k	$(s_k/C_k).(N_k/N)=f_k$	$n_k = nf_k$
	$\Sigma N_j = N$		$\Sigma n_j = n$

4.9 Cálculo del tamaño de muestra n.

4.9.1 Proporcional a la varianza. Se debe asignar un estimado del error (D) que esté asociado a la estimación de $\bar{Y}$.

$$n = \frac{N\ \Sigma N_j s_j^2}{N^2 D^2 + \Sigma N_j s_j^2}$$

4.9.2 Proporcional al costo.

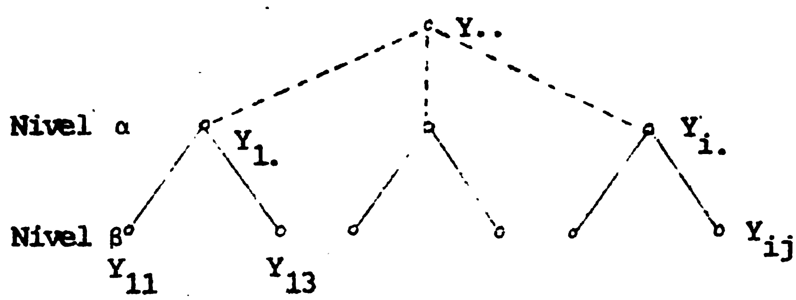
$$n = \frac{(\Sigma N s\sqrt{C_j})\,(\Sigma N_j s_j/\sqrt{C_j})}{N^2 D^2 + \Sigma N_j s_j^2}$$

5. MUESTREO JERARQUICO.

5.1 Usos. Se utiliza en los casos en que la fuente de variación se puede agrupar en clase jerárquicas o en algunos casos en que las unidades observacionales en la población se encuentran agrupadas en conglomerados. Se trata de una muestra no formada por individuos pero sí formada por conglomerados. Algunas de sus características sobresalientes son:

- Cada fuente de variación (jerárquica), está asociada con un componente de varianza σ_i^2.
- Es muy eficiente, cuando los niveles de jerarquía o conglomerados están perfectamente definidos.
- Existen tantas aleatorizaciones como etapas jerarquías o conglomerados existen.

Ejemplo 5.1 Polietápico de 2 jerarquías.



$$Y_{ij} = \mu + \alpha_i + \beta_j + n_{ij}$$

Donde:

Y_{ij} = variable respuesta

μ = media común

α_i = efecto de la jerarquía α

β_j = efecto de la jerarquía β

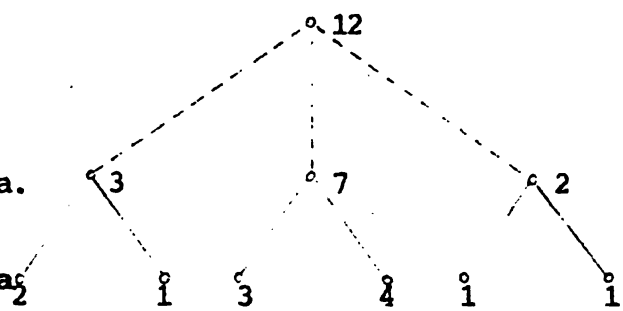
η_{ij} = error

El cálculo de estimadores se realiza mediante el análisis de varianza jerárquico. Se presentan los casos siguientes:

Ejemplo 5.2 Polietápico de 2 jerarquías y con igual número de observaciones

Etapa 1a.

Etapa 2a.



$n_1 = 3$

$n_2 = 2$

Cuadro 10. Análisis de varianza jerárquico con 2 etapas

F.V.	G.L.	S.C.	C.M.	E(CM)
Primera	$(n_1 - 1)$	$\Sigma Y^2_{i.}/n_2 - Y^2_{..}/n_1n_2$	CM_1	$\sigma^2_2 + n_2\sigma^2_1$
Segunda/Primera	$(n_1(n_2-1)$	$\Sigma\Sigma Y^2_{ij} - \Sigma Y^2_{i.}/n_2$	CM_2	σ^2_2
Total	$n_1n_2 - 1$	$\Sigma\Sigma Y^2_{ij} - Y^2_{..}/n_1n_2$		

$$Y = \Sigma\Sigma Y_{ij}/n_1n_2$$

$$s_{\bar{y}} = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_1n_2}}$$

$$SC\ P = \frac{3^2 + 7^2 + 2^2}{2} - \frac{12^2}{6} = 31 - 24 = 7$$

$$SC\ S/P = 2^2 + \ldots\ldots + 1^2 - \frac{3^2 + 7^2 + 2^2}{2} = 32 - 31\ 1$$

$$SC\ Tot. = 2^2 + \ldots\ldots + 1^2 - \frac{12^2}{6} = 32 - 24 = 8$$

Cuadro 11. Análisis de varianza y componentes de varianza

F.V.	G.L.	S.C.	C.M.	E(CM)
P	2	7	3.50	$\sigma_2^2 + 2\sigma_2^2$
S/P	3	1	0.33	σ_2
Total	5	8		

Estimación de los componentes de varianza. Partiendo del supuesto de que el cuadrado medio es un estimador del componente de varianza.

$$CM_1 = E(CM_1)$$
$$CM_2 = E(CM_2)$$

es decir:

$$\sigma_2^2 = CM_2 \qquad = 0.33$$

$$\sigma_1^2 = \frac{CM_1 - CM_2}{n_1} = 1.59$$

Variación total = $\sigma_1^2 + \sigma_2^2 = 1.59 + 0.33 = 1.92$

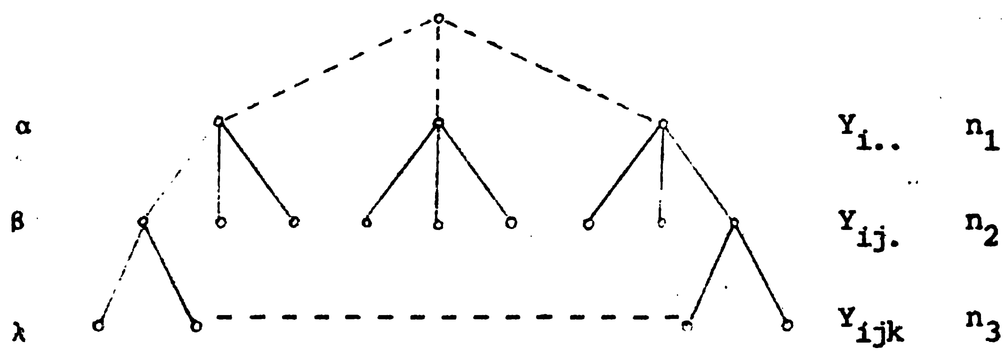
Variac. relativa= $\dfrac{\sigma_1^2}{\text{Total}} \cdot 100 = 83\%$

$$\frac{\sigma_2^2}{\text{Total}} \cdot 100 = 17\%$$

$$s_{\bar{y}} = \frac{1.59}{3} + \frac{0.33}{6} = 0.76$$

5.3 Análisis de varianza jerárquico con igual número de observaciones por etapa.

Ejemplo 5.3 Polietápico de 3 jerarquías, observaciones iguales.



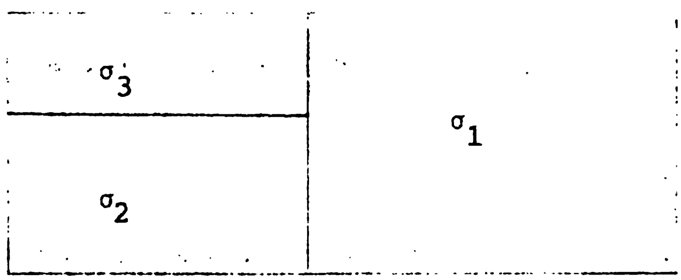
Cuadro 12. Análisis de varianza jerárquico con 3 etapas

F.V.	G.L.	S.C.
Primera	$(n_1 - 1)$	$\Sigma Y_{i..}^2/n_2n_3 - Y_{...}^2/n_1n_2n_3$
Segunda/Primera	$n_1(n_2 - 1)$	$\Sigma\Sigma Y_{ij.}^2/n_3 - \Sigma Y_{i..}^2/n_2n_3$
Tercera/Segunda	$n_1n_2(n_3 - 1)$	$\Sigma\Sigma\Sigma Y_{ijk}^2 - \Sigma\Sigma Y_{ij.}^2/n_3$
Total	$n_1n_2n_3 - 1$	$\Sigma\Sigma\Sigma Y_{ijk}^2 - Y_{...}^2/n_1n_2n_3$

Continuación del Cuadro 12.

F.V.	C.M.	E(CM)
Primera	CM_1	$\sigma_3^2 + n_3\sigma_2^2 + n_2n_3\sigma_1^2$
Segunda/Primera	CM_2	$\sigma_3^2 + n_3\sigma_2^2$
Tercera/Segunda	CM_3	σ_3^2

La variabilidad intrínseca de las etapas es diferente, pero se espera una categorización en función de su magnitud, tal que: varibilidad primera etapa > segunda etapa > tercera etapa, graficamente:



Estimación de componentes de varianza. Se parte del supuesto de que los cuadrados medios estiman la esperanza matemática de cada una de las etapas. Por lo que se pueden escribir las igualdades:

$$CM_1 = E(CM_1)$$
$$CM_2 = E(CM_2)$$
$$CM_3 = E(CM_3)$$

De la tercera igualdad se despeja $\hat{\delta}_3^2$ de la forma siguiente:

$$\hat{\sigma}_3^2 = CM_3 \qquad |4|$$

De la segunda y tercera igualdades se despeja $\hat{\sigma}_2^2$ por ecuaciones simultáneas.

$$\begin{array}{r} CM_2 = \sigma_3^2 + n_3\sigma_2^2 \\ - CM_3 = -\sigma_3^2 \\ \hline CM_2 - CM_3 = n_3\sigma_2^2 \end{array}$$

$$\hat{\sigma}_2^2 = \frac{CM_2 - CM_3}{n_3} \qquad |5|$$

De la primera y segunda igualdad en forma similar, se despeja $\hat{\sigma}_1^2$.

$$\hat{\sigma}_1^2 = \frac{CM_1 - CM_2}{n_2 n_3} \qquad |6|$$

Interpretación. Se realiza en términos relativos a la variación total; es decir, se totaliza sobre |4| |5| y |6| y se calculan los porcentajes para cada jerarquía o etapa.

$$\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2 = \text{Total}$$

$$\frac{\hat{\sigma}_1^2}{\text{Total}} \times 100 \quad = \text{Variabilidad relativa de la 1a. etapa}$$

$$\frac{\hat{\sigma}_2^2}{\text{Total}} \times 100 \quad = \text{Variabilidad relativa de la 2a. etapa}$$

$$\frac{\hat{\sigma}_3^2}{\text{Total}} \times 100 \quad = \text{Variabilidad relativa de la 3a. etapa}$$

El estimador de la media se càlcula en forma usual.

$$\bar{y} = \frac{\Sigma\Sigma\Sigma Y_{ijk}}{n_1 n_2 n_3}$$

La varianza de la media tiene 3 componentes para esta situación particular.

$$s_{\bar{y}}^2 = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_1 n_2} + \frac{\hat{\sigma}_3^2}{n_1 n_2 n_3}$$

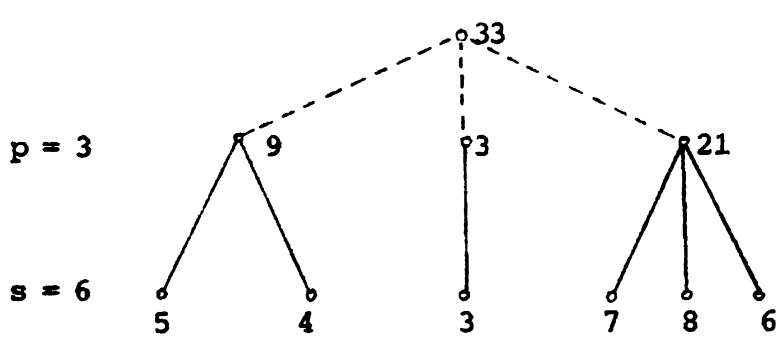
Cálculo de tamaño de muestra en cada etapa.

Se calcula con base en el costo (C_k) y la variabilidad ($\hat{\sigma}_k$) por jerarquía de acuerdo a la relación:

$$\hat{n}_k = \frac{\hat{\sigma}_k}{\hat{\sigma}_{k-1}} \sqrt{\frac{C_{k-1}}{C_k}}$$

5.4 Análisis de varianza jerárquico con desigual número de observaciones por etapa.

Ejemplo 5.4 Polietápico de 2 niveles.



Cuadro 13. Análisis de varianza jerárquico desigual número de observaciones, 2 etapas.

F.V.	S.C.	C.M.	E(CM)
p	$\Sigma Y_{i.}^2/n_{i.} - Y_{..}^2/n_{..}$	$(p - 1)$	$\sigma_2^2 + k\sigma_1^2$
s/p	$\Sigma\Sigma Y_{ij}^2 - \Sigma Y_{i.}^2/n_{i.}$	$(s - p)$	σ_2^2
Total	$\Sigma\Sigma Y_{ij}^2 - Y_{..}^2/n_{..}$	$(s - 1)$	

$$K = (\Sigma n_{i.} - \Sigma n_{i.}^2/\Sigma n_{i.})/(p - 1)$$

$$\bar{y} = \Sigma\Sigma Y_{ij}/s$$

$$s_{\bar{y}} = \sqrt{\hat{\sigma}_1^2/p + \hat{\sigma}_2^2/s}$$

$$\text{SC P} = \frac{9^2}{2} + 3^2 + \frac{21^2}{3} - \frac{33^2}{6} = 196.5 - 181.5 = 15$$

$$\text{SC S/P} = 5^2 + \ldots\ldots + 6^2 - \frac{9^2}{2} + 3^2 + \frac{21^2}{3} = 199 - 196.5 = 2.5$$

$$\text{SC Total} = 5^2 + \ldots\ldots + 6^2 - \frac{33^2}{6} = 17.5$$

$$k = (6 - \frac{14}{6})/2 = 1.8333$$

Cuadro 14. Análisis de varianza y componentes de varianza, 2 etapas

F.V.	S.C.	G.L.	C.M.	E(CM)	V.R. (%)
P	15.0	2	7.5	$\sigma_2^2 + 1.8333\ \sigma_1^2$	81.36
S/P	2.5	3	0.8333	σ_2^2	18.64
Total	17.5	5			

$$\hat{\sigma}_2^2 = 0.8333$$

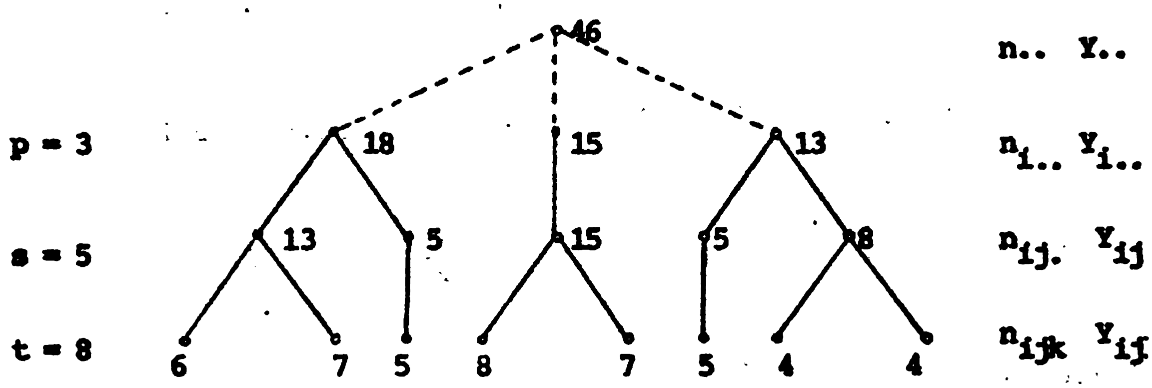
$$\hat{\sigma}_1^2 = \frac{7.5 - 0.8333}{1.8333} = 3.6364$$

$$\hat{\sigma}_1^2 + \hat{\sigma}_2^2 = 0.8333 + 3.64 = 4.4697$$

$$\bar{y} = \frac{33}{6} = 5.5$$

$$s_{\bar{y}} = \sqrt{\frac{3.6364}{3} + \frac{0.8333}{6}} = 1.1623$$

Ejemplo 5.5 Polietápico de 3 niveles desigual número de observaciones



Cuadro 15. Análisis de varianza jerárquico desigual número de observaciones, 3 etapas.

F.V.	S.C.	G.L.	E(CM)
P	$\Sigma Y_{i..}^2/n_{i..} - Y_{...}^2/n_{...}$	p - 1	$\sigma_3^2 + k_2\sigma_2^2 + k_1\sigma_2^2$
S/P	$\Sigma\Sigma Y_{ij.}^2/n_{ij.} - \Sigma Y_{i..}^2/n_{i..}$	s - p	$\sigma_3^2 \quad k_3\sigma_2^2$
T/S	$\Sigma\Sigma\Sigma Y_{ijk}^2/n_{ijk} - \Sigma\Sigma Y_{ij.}^2/n_{ij.}$	t - s	σ_3^2
Total	$\Sigma\Sigma\Sigma Y_{ijk}^2/n_{ijk} - Y_{...}^2/n_{...}$		

$$k_1 = (n_{...} - \Sigma n^2_{i..}/n_{...})/(p - 1)$$

$$k_2 = (\Sigma\Sigma n^2_{ij.}/n_{i..} - \Sigma\Sigma n^2_{ij.}/n_{...})/(p - 1)$$

$$k_3 = (n_{...} - \Sigma\Sigma n^2_{ij.}/n_{i..})/(s - p)$$

$$\bar{y} = \Sigma\Sigma\Sigma Y_{ijk}/t$$

$$s^2_{\bar{y}} = \hat{\sigma}^2_1/p + \hat{\sigma}^2_2/s + \hat{\sigma}^2_3/t$$

Cálculo aritmético.

$$SC\ P = \frac{18^2}{3} + \frac{15^2}{2} + \frac{13^2}{3} - \frac{46^2}{8} = 276.8333 - 264.5 = 12.3333$$

$$SC\ S/P = \frac{13^2}{} + \dots + \frac{8^2}{3} - \frac{18^2}{3} + \frac{12^2}{3} = 279 - 276.8333 = 2.1667$$

$$SC\ T/S = 6^2 + \dots\dots + 4^2 - \frac{13^2}{2} + \dots\dots + \frac{8^2}{2} = 280 - 279 = 1.0$$

$$SC\ Total = 6^2 + \dots\dots + 4^2 - \frac{46^2}{8} = 280 - 264.5 = 15.50$$

$$k_1 = (8 - \frac{3^2 + 2^2 + 3^2}{8})/2 = 2.625$$

$$k_2 = (\frac{2^2 + 1^2}{3} + \frac{2^2}{2} + \frac{1^2 + 2^2}{3} - \frac{2^2 + 1^2 + 2^2 + 1^2 + 2^2}{8})/2 = 1.7917$$

$$k_3 = (8 - \frac{2^2 + 1^2}{3}\ \frac{2^2}{2}\ \frac{1^2 + 2^2}{3})/2 = 1.3333$$

Cuadro 16. Análisis de varianza y componentes, 3 etapas

F.V.	S.C.	G.L.	C.M.	E(C.M.)	%
P	12.3333	2	6.1666	$\sigma_3^2 + 1.7917\sigma_2^2 + 2.625\sigma_1^2$	67.24
S/P	2.1667	2	1.0833	$\sigma_3^2 + 1.3333\sigma_2^2$	20.57
T/S/T	1.0000	3	0.3333	σ_3^2	
Total	15.5000	7			

$$\hat{\sigma}_3^2 = 0.3333$$

$$\hat{\sigma}_2^2 = \frac{1.0833 - 0.3333}{1.3333} = 0.5625$$

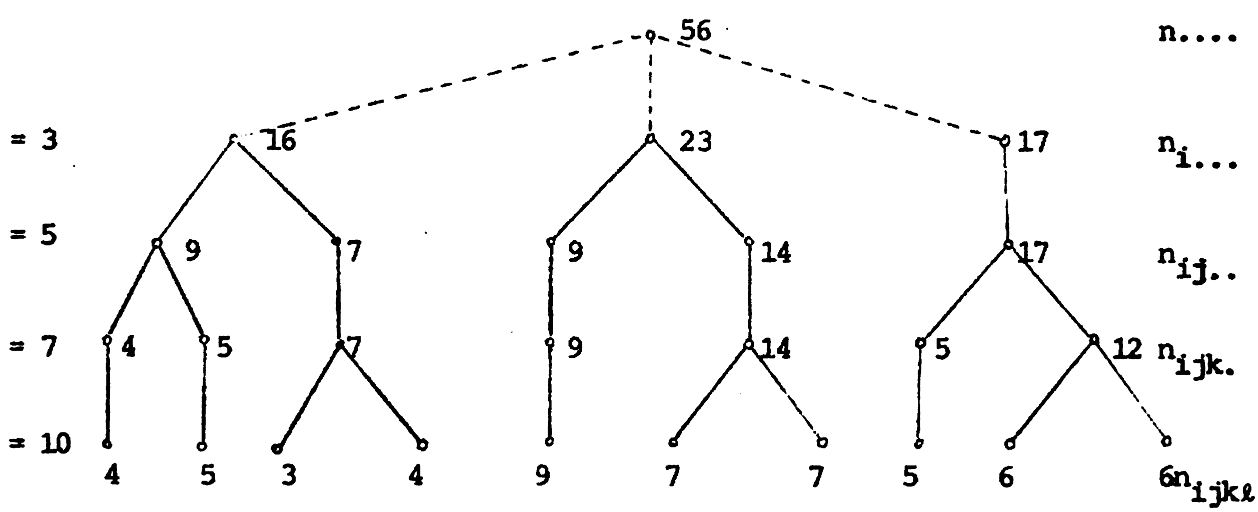
$$\hat{\sigma}_1^2 = \frac{6.1666 - 1.0833 - 0.2579}{2.625} = 1.8382$$

$$\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2 = 2.734$$

$$\bar{y} = 5.75$$

$$s_{\bar{y}} = \sqrt{\frac{1.8382}{} + \frac{0.5625}{} + \frac{0.3333}{}} = 0.8757$$

Ejemplo 5.6 Polietápico de 4 niveles



Cuadro 17. Análisis de varianza jerárquico de 4 etapas

F.V.	S.C.	G.L.	E(CM)
P	$\Sigma Y_{i...}^2/n_{i...} - Y_{....}^2/n_{....}$	p-1	$\sigma_4^2+k_4\sigma_3^2+k_2\sigma_2^2+k_1\sigma_1^2$
S/P	$\Sigma\Sigma Y_{ij..}^2/n_{ij..} - \Sigma Y_{i...}^2/n_{i...}$	s-p	$\sigma_4^2+k_5\sigma_3^2+k_3\sigma_2^2$
T/S/P	$\Sigma\Sigma\Sigma Y_{ijk.}^2/n_{ijk.}-\Sigma\Sigma Y_{ij..}^2/n_{ij..}$	t-s	$\sigma_4^2+k_6\sigma_3^2$
W/T/S/P	$\Sigma\Sigma\Sigma\Sigma Y_{ijk\ell}^2 - \Sigma\Sigma\Sigma Y_{ijk.}^2/n_{ijk.}$	w-t	σ_4^2
Total	$\Sigma\Sigma\Sigma\Sigma Y_{ijk\ell}^2 - Y_{....}^2/n_{....}$	w-1	

$$k_1 = (n_{...} - \Sigma n_{i..}^2/n_{...})/(p - 1)$$

$$k_2 = (\Sigma\Sigma n_{ij.}^2/n_{i..} - \Sigma\Sigma n_{ij.}^2/n_{...})/(p - 1)$$

$$k_3 = (n_{...} - \Sigma\Sigma n_{ij.}^2/n_{i..})/(s - p)$$

$$k_4 = (\Sigma\Sigma\Sigma n_{ijk}^2/n_{i..} - \Sigma\Sigma\Sigma n_{ijk}^2/n_{...})/(s - p)$$

$$k_5 = (\Sigma\Sigma\Sigma n_{ijk}^2/n_{ij.} - \Sigma\Sigma\Sigma n_{ijk}^2/n_{i..})/(s - p)$$

$$k_6 = (n - \Sigma\Sigma\Sigma n_{ijk}^2/n_{ij.})/(t - s)$$

$$\bar{y} = \Sigma\Sigma\Sigma Y_{ijk\ell}/w$$

$$s_{\bar{y}} = \sqrt{\frac{\hat{\sigma}_1^2}{P} + \frac{\hat{\sigma}_2^2}{S} + \frac{\hat{\sigma}_3^2}{t} + \frac{\hat{\sigma}_4^2}{W}}$$

Cálculo aritmético.

$$SC\ P = \frac{16^2}{4} + \ldots + \frac{17^2}{3} - \frac{56^2}{10} = 336.6666 - 313.6 = 23.0666$$

$$SC\ S/P = \frac{9^2}{2} + \ldots. + \frac{17^2}{3} - \frac{16^2}{4} + \ldots + \frac{17^2}{3} = 340.3333 - 336.666$$

$$= 3.6667$$

$$SC\ T/S = 4^2 + \ldots. + \frac{12^2}{2} - \frac{9^2}{2} + \ldots + \frac{17^2}{3} = 341.5 - 340.3333$$

$$= 1.1667$$

$$SC\ W/T/S = 4^2 + \ldots. + 6^2 - 4^2 + \ldots + \frac{12^2}{2} = 342.0 - 341.5 = 0.5$$

$$SC\ TOTAL = 4^2 + \ldots. + 6^2 - \frac{56^2}{10} = 342.0 - 313.6 = 28.4$$

$$k_1 = \left(10 - \frac{4^2 + 3^2 + 2^2}{10}\right)/2 = 3.55$$

$$k_2 = \left(\frac{2^2 + 2^2}{4} + \frac{1^2 + 2^2}{3} + \frac{3^2}{3} - \frac{2^2 + \ldots + 2^2}{10}\right)/2 = 2.2333$$

$$k_3 = \left(10 - \frac{2^2 + 2^2}{4} + \ldots\ldots + \frac{3^2}{3}\right)/2 = 1.6667$$

$$k_4 = \left(\frac{1^2 + 1^2 + 2^2}{4} + \ldots + \frac{1^2 + 2^2}{3} - \frac{1^2 + \ldots. + 2^2}{10}\right)/2$$

$$= 1.6167$$

$$k_5 = \left(\frac{1^2 + 1^2}{2} + \ldots\ldots + \frac{1^2 + 2^2}{3} - \frac{1^2 + 1^2 + 2^2}{4} + \ldots\ldots + \frac{1^2 + 2^2}{3}\right)/2$$

$$= 1.4168$$

$$k_6 = \left(10 - \frac{1^2 + 1^2}{2} + \ldots\ldots + \frac{1^2 + 2^2}{3}\right)/2 = 1.1667$$

Cuadro 18. Análisis de varianza y componentes de varianza, 4 etapas

F.V.	S.C.	G.L.	C.M.	E(CM)
P	23.0666	2	11.5333	$\sigma_4^2+1.6167\sigma_3^2+2.2333\sigma_2^2 + 3.55\sigma^2{}_1$
S/P	3.6667	2	1.8333	$\sigma_4^2+1.4168\sigma_3^2+1.6667\sigma_2^2$
T/S/P	1.1667	2	0.5833	$\sigma_4^2+1.1667\sigma_3^2$
W/T/S/P	0.5000	3	0.1666	σ_4^2
Total	28.4000	9		

$$\hat{\sigma} = 0.1666$$

$$\hat{\sigma}_3^2 = \frac{0.5833 - 0.1666}{1.1667} = 0.3572$$

$$\hat{\sigma}_2^2 = \frac{1.8333 - 0.5833 - 0.0894}{1.6667} = 0.6963$$

$$\hat{\sigma}1^2 = \frac{11.5333 - 1.8333 - 0.4659}{3.55} = 2.6012$$

$$\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2 + \hat{\sigma}_4^2 = 3.8213$$

% de la variación total.

P = 68.07

S/P = 18.22

T/S = 9.35

W/T = 4.36

TOTAL = 100.00

$\bar{y} = 5.6$

$$s_{\bar{y}} = \sqrt{\frac{2.6012}{3} + \frac{0.6963}{5} + \frac{0.3572}{7} + \frac{0.1666}{10}} = 1.0363$$

Si existen más de 4 jerarquías en el diseño de muestreo, los coeficientes k_i para evaluar los componentes de varianza, se obtienen por extensión de las matrices siguientes:

$$\begin{bmatrix} k_2 & k_1 \\ k_3 & 0 \end{bmatrix} \qquad \begin{bmatrix} \Sigma\Sigma n_{ij}^2 \; f_i & \Sigma n_{i.}^2 \; f_i \\ \Sigma\Sigma n_{ij}^2 \; f_{ij} & 0 \end{bmatrix}$$

Donde:

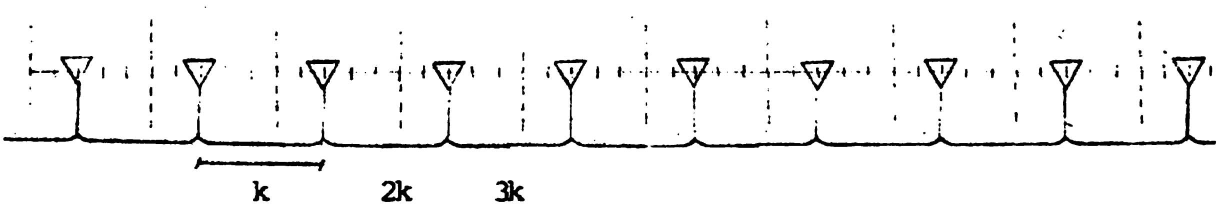
f_i es la media geométrica $(1/n_{i.} - 1/n_{..})/(p - 1)$

f_{ij} es la media geométrica $(1/n_{ij} - 1/n_{i.})/(s - p)$

6. MUESTREO SISTEMATICO

6.1 Usos. Se utiliza en los casos con intervalo de muestreo fijo, constante (k) o cuando la cantidad de trabajo necesario para extraer una muestra al azar puede ser bastante considerable es decir, el número de unidades a relacionar es muy alto; por ejemplo, para conseguir una muestra con una presión de selección del 5% sobre 300000 elementos, sería necesario seleccionar 15000 números de la tabla de números al azar, es más aconsejable sortear una sola vez un elemento de los 5 primeros y luego continuar con dada quinto elemento hasta completar 15000; es decir, tomar uno de cada cinco. Se define previamente el número de zonas, para seleccionar aleatoriamente un elemento en la primera zona, luego se continua con la enumeración de elementos a distancia o intervalo (k) constante, hasta completar los n elementos de la muestra, como en la ilustración siguiente:

Selección aleatoria



6.2 Estimadores. El cálculo aritmético de los estimadores se realiza, teniendo en cuenta el marco de referencia. Si los individuos Y_i no conllevan un ordenamiento o arreglo de los datos, se utilizan las fórmulas definidas en el muestreo irrestricto al azar, es decir, Promedio aritmético

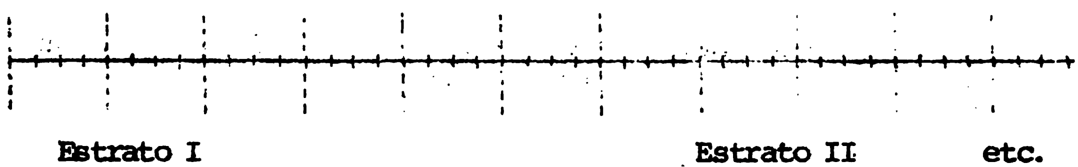
$$\bar{y} = \frac{1}{n}\,\Sigma Y_i$$

Varianza

$$s^2 = \frac{1}{n}\,(\Sigma Y_i - \bar{y})^2$$

$$s_{\bar{y}} = \frac{s}{n} \cdot \frac{N-n}{N-1}$$

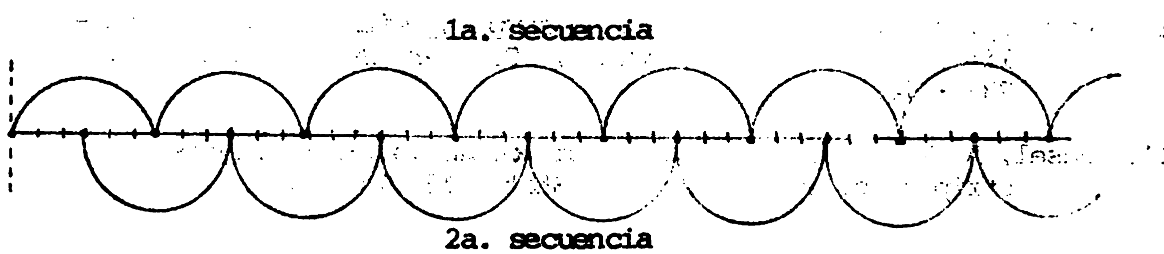
En cambio si las unidades adyacentes son más parecidas entre sí que las unidades más distantes, se sugiere estratificar las observaciones Y_i.



$$\bar{y}_{estrato} = \Sigma \frac{N_i^2}{N}\,\bar{y}_i$$

$$s^2_{\bar{y}estrato} = \Sigma \frac{N_i^2}{N}\,s^2_{\bar{Y}_i}$$

Para ambas suposiciones se recomienda por lo menos 2 secuencias sistemáticas, ejemplo.



REFERENCIAS

1. Avadhani, M. S., and Sukhatme, B. V. 1965. Controlled simple random sampling. F. Ind. Soc. Agri. Stat., 17, 34 - 42.

2. Bryant, E. C., Hartley, H. O. and Jessen, R. J. 1960. Design and estimation in twoway stratification. F. Amer. Stat. Assoc., 55, 105 - 124.

3. Cochran, W. G. 1963. Sampling techniques, second edition, John Wiley and Sons, Inc., New York, London.

4. Evans, W. D. 1951. On stratification and optimum allocations, F. Amer. Stat. Assoc., 46, 95 - 104.

5. Fellegi, I. P. 1963. Sampling with varying probabilities without replacement: Rotating and non-rotating samples, F. Amer. Stat. Assoc., 58, 183 - 201.

6. Fisher, R. A. and Yates, F. 1953. Statistical tables for biolo gical, agricultural, and medical research, 4th edition. Oliver and Boyd, Ltd.

7. Gautschi, W. 1957. Some remarks on systematic sampling. Ann. Math. Stat., 28, 385 - 394.

8. Hansen, M. H., Hurwitz, W. N., and Madow, W. G. 1953. Sample survey methods and theory, Vol I and II, New York: John Wiley and Sons, Inc.

9. __________., and Hurwitz, W. N. 1943. On the theory of sampling from finite populations, Ann. Math. Stat., 14, 333 - 362.

10. Hartley, H. O. 1966. Systematic sampling with unequal probability and without replacement, F. Amer. Stat. Assoc., 61, 739 - 748.

11. Hasel, A. A. 1942. Estimation of volume in timber stands by strip sampling. Ann. Math. Stat., 13, 179 - 206.

12. Hoel, P. G. 1954. Introduction to mathematical statistics, 2nd edition. John Wiley & Sons, Inc., New York.

13. Houseman, E. E. 1976. El muestreo por áreas en la agricultura. Statistical Reporting service. United States Department of Agriculture, Washington, D.C. U.S.A.

14. Kish, L. 1965. Survey sampling. John Wiley and Sons, New York.

15. Kitagawa, T. 1956. Some contributions to the design of sample survey, Sankhya, 17, 1 - 36.

16. Madow, L. H. 1946. Systematic sampling and its relation to other sampling designs. F. Amer. Stat. Assoc., 41, 204-217.

17. Madow, W. G., and L. H. 1944. On the theory os systematic sampling. Ann. Math. Stat., 15, 1 - 24.

18. Mahalanobis, P. C. 1952. Some aspects of the design of sample surveys, Sankhya, 12, 1 - 7.

19. Quiroga, V. Manual práctico para análisis de experimentos de campo. IICA. Publicaciones misceláneas No. 142. 1976. 113 p.

20. __________. Manual de estadística descriptiva. IICA. Publicaciones misceláneas No. 147. 1977. 82 p.

21. __________. Manual para estimar parámetros de seis modelos aplicados a fenómenos sociales, económicos y biológicos. IICA. Publicaciones misceláneas No. 145. 1977. 36 p.

22. Rao, J. N. K, Hartley, H. O. and Cochran, W. G. 1962. On a simple procedure of unequal probability sampling without replacement. F. Roy Stat. Soc., Series B, 24, 482 - 491.

23. Sampford, M. R. 1962. Methods of cluster sampling with and without replacement for clusters of unequal sizes, Biometrika, 49, 27 - 40.

24. Sampford, M. R. 1962. An introduction to sampling theory. Oli ver and Boyd. Edinburgh and London.

25. Snedecor, G. W. and Cochran, W. G. Statistical methods. Sixth edition. Iowa State. Ames, Iowa, U.S.A.

26. Sukhatme, B. V. and Koshal, R. S. 1959. A contribution to double sampling, F. Ind. Soc. Agri. Stat., 11, 128 - 144.

27. Sukhatme, P. V. and Panse, V. G. 1951. Crop surveys in India, F. Ind. Soc. Agri. Stat., 3. 97 - 168.

28. ____________. and Balkrisha. V. Sampling theory of surveys with applications. Iowa State University Press, Ames, Iowa, U.S.A.

29. ____________. 1935. Contribution to the theory of the representative method, F. Roy. Stat. Soc. Suppl., 2, 253 - 268.

30. ____________. 1956. Teoría de encuestas por muestreo con aplicaciones. Fondo de cultura económica, Méjico, D.F.

31. Steel, R. G. D. and Torrie, J. H. 1960. Principles and procedures of statistics. New York, Toronto, London.

32. Stuart A. 1962. Basic ideas of scientific sampling. Charles Griffin, London.

33. U.S. Bureau of the Census: Curso suplementario para un estudio de censo sobre encuestas y censos, conferencias sobre muestreo, isp Supplemental Course Series, No. 1. (Versión en español) Washington, D.C., 1971.

34. Weatherburn, C. E. 1947. A first course in mathematical statis tics. Cambridge University Press.

35. Yates, F. 1960. Sampling methods for censuses and surveys, 3rd ed. Charles Griffin, London.

36. _______. 1953. Sampling methods dor censuses and surveys, end edition. Charles Griffin and Co., Ltd., London

37. Zarcovich, S. S. Sampling methods and census. FAO, Rome. 1965.

AGRICOLAS

E
es
es
bi

S
bi
H
na

La dirección postal de la Sed
Cable: IICASANJOSE; Telex:

* En proceso de ingreso.

El Programa Cooperativo Regional para ... o, Centroamérica y Panamá PROMECAFE está ... etarías de Estado, los Organismos del Café de los ... estigación y Enseñanza, CATIE y el Organismo Interna...

PROMECAFE fue creado el 31 de ener... Operaciones en San José, Costa Rica.