Muestreo

Mucho de las acciones y decisiones que se toman están basados en la información de una muestra.

La pregunta que siempre se hace, es:

- ✓ ¿qué tamaño de muestra es suficiente para obtener una buena aproximación de la población de estudio?,
- ✓ ¿Qué estimaciones se debe realizar con la muestra?,
- ✓ ¿Qué método de muestreo se debe realizar?

Existen muchas razones para el uso del muestreo:

- ✓ Costo reducido
- ✓ mayor rapidez de la información
- ✓ mayores posibilidades de realizar
- ✓ mayor exactitud en las mediciones

En muchos aspectos, mayormente en lo social y económico, la encuesta es la manera de adquirir información para el estudio;

La opinión pública sigue siendo un factor importante para los periódicos, estas encuesta por muestreo pueden ser clasificadas como descriptivas o analíticas; lo descriptivo para obtener información de grandes grupos y lo analítico para el estudio comparativo de grupos.

Una encuesta por muestreo contempla los siguientes aspectos:

- ✓ Exposición clara de los objetivos
- ✓ Definición de la población bajo muestreo
- ✓ Que datos se debe colectar
- ✓ Que nivel de precisión se desea.
- ✓ Que métodos de medición e inspección de la población
- ✓ Cual es el marco, totalidad de las unidades de muestreo que hace la población.
- ✓ Como seleccionar la muestra
- ✓ Como utilizar una encuesta piloto
- ✓ Como organizar el trabajo en campo
- ✓ Formas para resumir la encuesta
- ✓ Que información es obtenida para encuestas futuras.

En el campo forestal, los estudios en base a muestreo son muy frecuentes, los objetivos también son descriptivos como analíticos, por ejemplo estudiar:

- ✓ la diversidad de especies de árboles en una localidad,
- ✓ la calidad de madera, evaluar rendimiento,
- ✓ el efecto de los componentes del suelo en el desarrollo de los árboles,
- ✓ el efecto de un manejo cultural en el rendimiento de cultivos asociados con árboles o con rotación de cultivos.

Los objetivos estadísticos son estimar la media, la variancia, la proporción, el total, la distribución, etc.

Muestreo Probabilístico.

Se entiende al proceso de obtener un conjunto de muestras, en las cuales cada elemento de la muestra tiene una probabilidad de ser escogida.

Muestreo simple aleatorio (al azar).- es un método de muestreo mas simple, en el cual se considera que cada elemento a ser considerada en la muestra tiene igual probabilidad de ser escogida.

Por ejemplo se dispone de 5 elementos que conforman la población { A, B, C, D y E} si se obtiene muestras de tamaño 3, el numero de muestras posibles seria:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \binom{5}{3} = \frac{5!}{3!2!} = 10$$

ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE

Cada una de las muestras tiene igual probabilidad de ser escogida.

La notación en muestreo es importante, las Letras mayúsculas definen la población y las minúsculas la muestra.

Total = Y; Promedio de Y= $\overline{Y} = \frac{\sum_{i}^{N} yi}{N}$;

en la muestra el total se expresa por 'y' y el promedio por $\overline{y}$

La media muestral es un estimador insesgado del promedio de la población.

$\hat{Y} = N\bar{y}$ es un estimador insesgado de la población total 'Y'.

Variancia de la población, $\sigma^2 = \frac{\sum_i^N (y_i - \bar{Y})^2}{N}$;

La variancia de la muestra se define por $s^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1}$, que estima a la variancia de la población.

La variancia de la media de una muestra simple al azar por

$$V(\bar{y}) = \frac{S^2}{n}(1-f)$$

Donde f = n/N.

Y el error estándar de la media es $S_{\bar{y}} = \frac{S}{\sqrt{n}}\sqrt{1-f}$

La variancia del estimador del total es $V(\hat{Y}) = \frac{N^2 S^2}{n}(1-f)$ y su error estándar $S_{\hat{y}} = N\frac{S}{\sqrt{n}}\sqrt{1-f}$

El factor (1-f) se denomina factor de corrección por población finita o corrección por finitud

<u>Estimación del tamaño de muestra con datos continuos.</u>

Si $\bar{y}$ es el promedio de una muestra simple aleatoria, entonces:

$$\Pr(\left|\bar{y} - \bar{Y}\right| \geq d) = \alpha$$

d = margen de error elegido y α es una probabilidad pequeña

Si $\bar{y}$ tiene una distribución normal, entonces el tamaño de muestra 'n' puede ser calculada como:

$$n = \frac{\left(\frac{zS}{d}\right)^2}{1 + \frac{1}{N}\left(\frac{zS}{d}\right)^2}$$

Una primera aproximación seria:

$$n0 = \left(\frac{zS}{d}\right)^2$$; significa que

$$n = \frac{n0}{1 + \frac{1}{N} n0}$$

El valor de Z es igual a 1.96 para 0.95; por lo general se aproxima a 2.

Ejemplo 1. Estimación del tamaño de muestra para la variable altura de 1000 árboles de capirona.

Bajo el supuesto de tener un estimado de la desviación estándar (aproximado con una muestra pequeña, o referencia de la población, por ejemplo S = 79.2 y un estimado de la magnitud de la media por ejemplo de 250, se puede estimar el tamaño de la muestra bajo una tolerancia y una probabilidad de error dada.

Por ejemplo una probabilidad de error del 0.05, con una tolerancia del 10% (error relativo), se tendría la siguiente estimación:

$Z = 1.96$ (probabilidad de error de 0.05)
$d / |250| = 0.10$ (10% de tolerancia respecto de la media)
$d = 25$

$$n0 = (1.96*76.2/25)^2 = 35.68$$
$$n = 35.68 / (1+ 35.68/1000) = 34.46 \approx 35$$

En algunas situaciones, la estimación puede ser en función del CV (coeficiente de variación), dado que CV = S/Media. Si se asume un 30% de CV entonces S/Media = 0.30 y una tolerancia del 10% (error relativo 0.10)

$$n0 = (1.96*30/10)^2 = 34.57$$

$$n = 34.57 / (1+ 34.57/1000) = 33.41 \approx 34$$

en esta segunda forma, no es necesario las estimaciones parciales de la desviación estándar y el promedio, puesto que el coeficiente de variación esta en función de ambos.

Para esta segunda forma, es necesario de todas maneras tener una estimación del CV aproximado con una muestra pequeña para estimar un tamaño de muestra optimo.

Muestreo para proporciones y porcentajes.

En algunos casos se requiere estimar la proporción o porcentaje o el total de elementos que tienen alguna característica, por ejemplo % de semillas infectadas en un lote para comercializar, % de plantas con cierta enfermedad en el campo, % de germinación de semilla certificada, Proporción de árboles de una especie que tenga un propósito, por ejemplo medicinal, % de tablillas de primera en árboles de una determinada procedencia, etc.

El total de unidades de la población es 'A' y el estimado de la muestra es 'a'; la proporción de unidades de A en la población es P=A/N y el estimador muestral es p = a/n.

En poblaciones infinitas, el trabajo estadístico es con la Binomial aplicado a 'p' y 'a'

En poblaciones finitas, la distribución correcta es la Hipergeometrica.

Variancia de los estimadores.

$$V(p) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right) \; ; \; Q = 1 - P$$

P y Q pueden ser estimados de la muestra, entonces la variancia estimada de p seria:

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

Para la estimación del total A , $\hat{A}$ es Np y la variancia:

$$V(\hat{A}) = \frac{N^2 PQ}{n}\left(\frac{N-n}{N-1}\right) \quad v(\hat{A}) = Npq\left(\frac{N-n}{n-1}\right)$$

Aproximación normal.- En algunos casos, para estimar limites de confianza, se puede utilizar la distribución normal para tener aproximaciones a estos limites.

$$p \pm \left[Z\sqrt{1-f}\,\sqrt{pq/(n-1)} + \frac{1}{2n} \right]$$

donde f=n/N; el factor 1/2n es la corrección por continuidad, Z es el desvío normal correspondiente a la probabilidad de confianza.

SI n es suficientemente grande, algunos prefieren utilizar son la expresión $\sqrt{pq/n}$ y no utilizar la corrección de 1/2n. Si N es suficientemente grande o infinito, entonces f se considera ‘0’.

Estimación de la muestra para proporciones

Las unidades de la población son clasificadas en dos grupos o clases. Se da como criterio algún margen de error ‘d’ en la proporción estimada ‘p’ de unidades perteneciente a una clase y hay un pequeño riesgo α que se dispone aceptar, es decir:

$$\Pr\left(|p-P| \geq d\right) = \alpha$$

Se supone un muestreo simple aleatorio y se considera que p se distribuye normalmente.

Una primera estimación es n0, dado por:

$$n0 = \frac{z^2 pq}{d^2}$$

Si n0/N no es despreciable, el numero de muestra se obtiene como:

$$n = \frac{n0}{1 + \frac{n0 - 1}{N}} = \frac{n0}{1 + \frac{n0}{N}}$$; para n0 considerable, mayor de 30.

La formula es valida si d, p y q están todas expresadas en porcentaje.

Ejemplo. Se tiene la evaluación de calidad de 1000 árboles. Según la estimación sobre la proporción de árboles de calidad, se estima un 77%. Se requiere obtener una muestra optima con una tolerancia del 10% y una confiabilidad del 95%.

Información del registro de árboles.

Árbol	Altura	Calidad
1	125	M
2	180	M
3	210	B
4	125	M
5	190	M
6	142	M
999	165	M

1000	250	B

Al 95% de confianza, se tiene el valor Z = 1.96
P=0.77, Q=0.23

Entonces: $n_0 = \dfrac{1.96^2 * 77 * 23}{10^2} = 68.03$

Como es una población finita entonces, $n = \dfrac{68.03}{1 + \dfrac{68.03}{1000}} = 63.69 \approx 64$

El tamaño de muestra optimo es 64.

<u>Muestreo sistemático</u>.- Si los elementos de una población están enumerados, es posible obtener con facilidad una muestra sistemática. Este método consiste en:

SI N es el tamaño de la población y se requiere una muestra de tamaño ‘n’, entonces se determina el valor de salto como k=N/n; dentro del primer salto, desde 1 hasta k, se obtiene un elemento al azar, los siguientes se extraen cada salto secuencialmente, hasta completar las ‘n’ muestras.

Por ejemplo, en la siguiente lista numerada: x1, x2,.....x200; se requiere una muestra n=15; entonces:

K=200/15 = 13.33 ~ 13

Entre los primeros 13 se extrae por ejemplo x8; entonces los siguientes elementos serán: X21 X34,..., X190; son en total 15 valores.

La variancia de la media de una muestra sistemática es:

$$V(\overline{y}_{sis}) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S_1^2$$

Donde $S_1^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{i.}\right)^2$

También conocido como el Cuadrado medio dentro de grupo.

La media de una muestra sistemática es mas precisa que la media de una muestra al azar simple si $S^2_1 > S^2$.

Muestreo estratificado

La población N esta dividido en subpoblaciones o estratos N1, N2, etc. Estas subpoblaciones no se superponen.

Para tener un beneficio de la estratificación, se debe conocer el tamaño de cada estrato Ni.

Este muestreo es bastante aplicado a poblaciones de personas, cuando se realiza encuestas por localidad; la localidad es el estrato.

En campo para medidas agronómicas si se tiene criterios de estratificación
Para fines de estudio, cada estrato es considerada como una pequeña población, y la estimación se realiza en cada estrato, para luego extenderla a la población.

Definida los estratos y sus tamaños, se procede a determinar el tamaño optimo de muestra para cada estrato.

N= total de la población
n = total de la muestra
Nh = total en el estrato h; Wh = Nh / N es la proporción del estrato
Para el calculo de los tamaños de muestra, se estima el tamaño n de muestra total en función de V.

V = (d/t)² , Los tamaños por estrato están dados por: d = margen de error o tolerancia y
t = valor tabular al α deseado.

Asignación Proporcional

$$n = \frac{\sum \frac{W_h^2 S_h^2}{w_h}}{V + \frac{\sum W_h^2 S_h^2}{N}} \qquad n_0 = \frac{\sum W_h s_h}{V} \qquad n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Tamaño por estrato

$$n_h = n \frac{W_h S_h}{\sum W_h S_h} = n \frac{N_h S_h}{\sum N_h S_h}$$

Encuestas

El estudio de poblaciones mediante encuesta son muy utilizados, porque en una consulta se solicita muchas preguntas, cada una es una variable de estudio. Esta técnica es ampliamente estudiada y requiere mucho conocimiento para elaborar y establecer la estrategia del levantamiento de la encuesta.

En esta parte solo se mencionara como determinar el numero de encuestas apropiado.
Se asume el criterio de la no-respuesta, método presentado por: Birnbaum y Sirken (1950)

Para un tolerancia (margen de error) d, el tamaño seria:
$n = \frac{Z_\alpha^2 PQ}{d^2}$; para P=Q=0.5, y α la probabilidad de error

Resulta: $n = \frac{Z_\alpha^2}{4d^2}$

	Confiabilidad		
Tolerancia	99	95	90
%	2.58	1.96	1.64
5	666	384	269
10	166	96	67
15	74	43	30
20	42	24	17
25	27	15	11