2.9 Determinación del tamaño de Muestras para Medias y Totales

Elementos a Considerar en el tamaño de muestra.
La pregunta con la que inicialmente se inician los trabajos de una encuesta es casi inevitablemente la referencia al tamaño de muestra necesario, sin embargo la respuesta no se puede dar a la ligera, pues requiere de la determinación de varios aspectos.

a) Tener al menos una idea aproximada de la magnitud del error permisible en la estimación.

b) Elegir el nivel de confianza, esto es, la probabilidad de ubicarse dentro del margen de error permisible.

c) Disponer información sobre la varianza de las principales variables que son objetivo de la encuesta el origen de esta información puede ser una prueba piloto, datos de encuestas similares aplicadas con anterioridad o incluso conjeturas sobre las distribuciones y varianzas asociadas a las variables de interés.

d) Plantear una función que involucre todos estos elementos para obtener el valor del tamaño de muestra n.

e) Debido a que las encuestas usualmente tienen muchas preguntas, se debe determinar cual ó cuales son las más importantes para que en base a ellas, se calculen tamaños opcionales de muestra.

f) Cuando se desean presentar resultados por subdivisiones de la población, se debe calcular por separado el tamaño de muestra para cada subdivisión y tomar el tamaño de muestra total como la suma de los valores de los tamaños calculados para las subdivisiones.

Error de Muestreo

La diferencia entre el valor del parámetro de la población y el valor que toma el estimador se denomina error de muestreo.

Error de Muestreo = $\theta - \hat{\theta}$

Como el error de muestreo está en términos de un parámetro desconocido, no es posible conocer este error.

Sin embargo, es posible establecer una relación probabilística en torno a un error máximo admisible d, de la siguiente manera siguiente.

$$P\left(\left|\overline{Y} - \overline{y}\right| > d\right) = \alpha$$

$$\therefore \quad P\left(\left|\overline{Y} - \overline{y}\right| \leq d\right) = 1 - \alpha$$

$$\therefore \quad P=\left\{\frac{\left|\overline{Y}-\bar{y}\right|}{\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}}\le\frac{d}{\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}}\right\}=1-\alpha$$

$$\therefore \quad P=\left\{-\frac{d}{\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}}\le\frac{\overline{Y}-\bar{y}}{\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}}\le\frac{d}{\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}}\right\}=1-\alpha$$

Si se considera el supuesto de normalidad para $\bar{y}$, se plantea la siguiente igualdad:

$$Z_{(1-\alpha/2)}=\frac{d}{\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}}$$

Donde $Z_{(1-\alpha/2)}$ es el valor tal que $P=\left(Z\le Z_{(1-\alpha/2)}\right)=\left(1-\alpha/2\right)$ si Z ≈ N(0,1). De aquí se llega a la siguiente expresión que define una varianza deseada por el investigador y que se define en el cociente del miembro izquierdo:

$$\frac{d^2}{Z^2_{(1-\alpha/2)}}=\left(1-\frac{n}{N}\right)\frac{s^2}{n}$$

De esta expresión se despeja n

$$d^2=Z^2_{(1-\alpha/2)}\left(1-\frac{n}{N}\right)\frac{s^2}{n}$$

$$Nnd^2=Z^2_{(1-\alpha/2)}(N-n)s^2$$

$$=Z^2_{(1-\alpha/2)}s^2N-Z^2_{(1-\alpha/2)}s^2n$$

$$\therefore$$

$$n=\frac{NZ^2_{(1-\alpha/2)}s^2}{Nd^2+Z^2_{(1-\alpha/2)}s^2}$$

Se dividen numerador y denominador entre Nd^2:

$$n=\frac{\frac{Z^2_{(1-\alpha/2)}s^2}{d^2}}{1+\frac{Z^2_{(1-\alpha/2)}s^2}{Nd^2}}$$

La expresión $\frac{Z^2_{(1-\alpha/2)}s^2}{d^2}$ corresponde al tamaño de muestra para una población muy grande o una selección con reemplazo y se identifica como $\mathbf{n_o}$ para sustituirla en la fórmula anterior.

$$n = \frac{n_o}{1+\frac{n_0}{N}}$$

$$n_o = \frac{Z^2_{(1-\alpha/2)}s^2}{d^2}$$

Note que $\mathbf{n < n_o}$ y que para N muy grande n converge a n_o.

Ejemplo 2.2

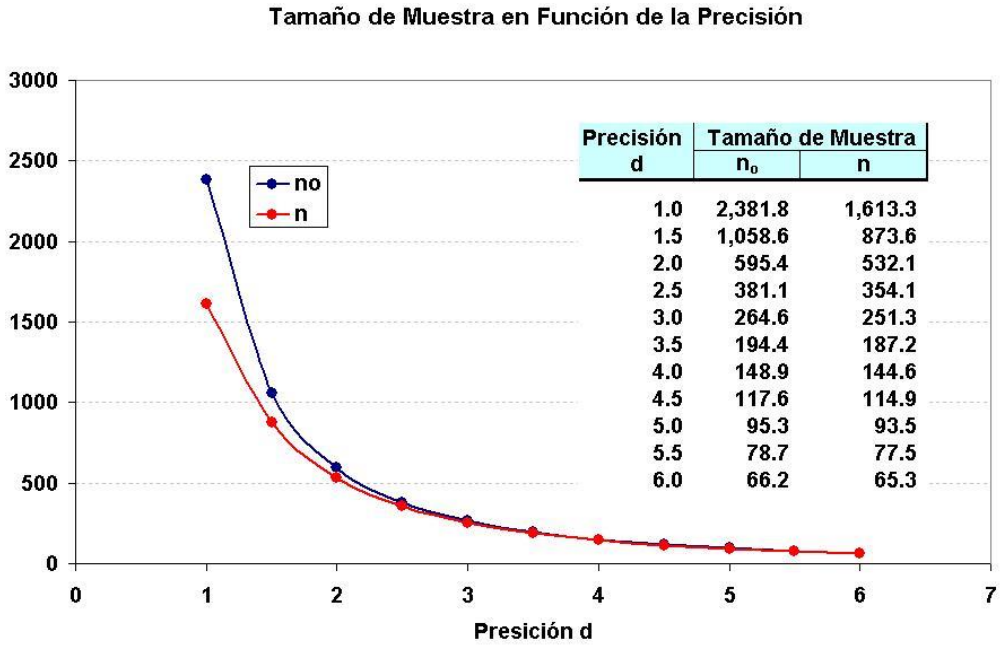
Se desea estimar el peso promedio de una población de 5000 cerdos con una precisión de d=2 kg. Se supone S^2= 380 y se desea una confianza de 95%. Calcule el tamaño de muestra necesario.

d = 5 = $Z_{(1-\alpha/2)}$ 1.96 N=5000 S^2= 380

$$n_o = \frac{Z^2_{(1-\alpha/2)}s^2}{d^2} = \frac{(1.96)^2(380)}{4} = 364.95 \qquad n = \frac{n_o}{1+\frac{n_0}{N}} = \frac{364.954}{1+\frac{365.954}{5000}} = 340.12$$

n = 341

El tamaño de muestra es particularmente sensible a la precisión. En la siguiente tabla se presentan tamaños de muestra en función de la precisión. Tamaño de la población **N** = 5000, varianza $\mathbf{S^2}$ = 620 y coeficiente de confianza de 95% **Z** = 1.96. Se ha variado la precisión **d** desde 1 hasta 6. El tamaño de **n** oscila entre 66 y 1614 y $\mathbf{n_o}$ entre 66 y 2382.

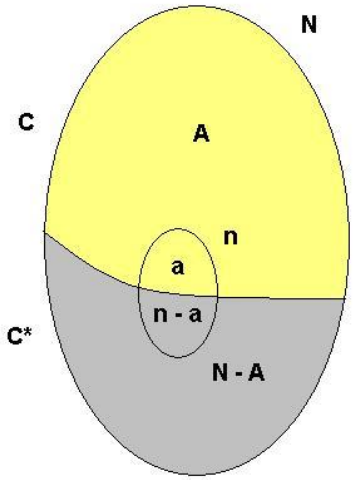


Precisión d	Tamaño de Muestra n_o	n
1.0	2,381.8	1,613.3
1.5	1,058.6	873.6
2.0	595.4	532.1
2.5	381.1	354.1
3.0	264.6	251.3
3.5	194.4	187.2
4.0	148.9	144.6
4.5	117.6	114.9
5.0	95.3	93.5
5.5	78.7	77.5
6.0	66.2	65.3

2.10 Muestreo Aleatorio Simple para Proporciones

Frecuentemente se desea estimar a proporción o el total de unidades en una población que poseen determinada característica o atributo.

Supóngase que la población consta de dos clases C y C*. Los elementos de interés pertenecen a la clase C. La población consta de N elementos de los cuales A son de la clase C y N-A del complemento C*, la muestra aleatoria simple tiene n elementos con **a** pertenecientes a la clase C. El parámetro de interés es la proporción definida por el cociente



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

$$Q = 1 - P$$

El estimador natural de la proporción poblacional P es la proporción muestral $p = \frac{a}{n}$

El estimador del total de elementos de la clase de interés A, es el producto $\hat{A} = Np$.

La estimación de proporciones se puede ver como un caso particular de la estimación de medias que involucran variables que adoptan valores 0 y 1 de acuerdo a la siguiente regla.:

$$\text{Sea } y_i = \begin{cases} 1 & \text{si la unidad} \in C \\ 0 & \text{si la unidad} \notin C \end{cases}$$

De aquí se concluye de manera inmediata que la proporción es equivalente a la media de y definida como variable dicotómica.

$$\overline{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{A}{N} = P$$

Además debido a la dicotomía de yi se tiene el siguiente resultado.

$$A = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} y_i^2 \quad \text{semejante para la muestra} \quad a = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i^2$$

La S^2 también adopta una forma particular en razón de la dicotomía.

$$\begin{aligned} S^2 &= \frac{1}{N-1}\left[\sum_{i=1}^{N}(y_i-\bar{Y})^2\right] \\ &= \frac{1}{N-1}\left[\sum_{i=1}^{N}y_i^2 - N\bar{Y}^2\right] \\ &= \frac{1}{N-1}\left[NP - NP^2\right] \\ &= \frac{N}{N-1}P(1-P) \end{aligned}$$

La varianza del estimador de la proporción adopta una forma alternativa:

$$\begin{aligned} V(p) &= \left(1-\frac{n}{N}\right)\frac{S^2}{n} \\ &= \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{N}{N-1}P(1-P) \end{aligned}$$

Al hacer las cancelaciones adecuadas se obtiene la fórmula de la varianza de p

$$\boxed{V(p)=\frac{N-n}{N-1}\frac{PQ}{n}}$$ y su error estándar $$\boxed{EE(p)=\sqrt{\frac{N-n}{N-1}\frac{PQ}{n}}}$$

De modo semejante al caso de $\bar{y}$ cuya varianza está en función de S^2 la cuál es desconocida la mayoría de las veces, se observa que la V(p) está en función precisamente del parámetro P.

En vista de ello lo que se hace es utilizar el estimador **p** para estimar su varianza.

El estimador insesgado de la S^2 en términos de p adopta la siguiente expresión :

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

Así al sustituir el estimador s^2 en la fórmula para la estimación de la varianza de la media, para p en nuestro caso se obtiene el estimador insesgado de la varianza de p.

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

Finalmente se tendrá la fórmula para un estimador insesgado: $$\boxed{\hat{V}(p)=\frac{N-n}{n-1}\frac{pq}{N}}$$

El error estándar estimado de p se calcula: $$\hat{\mathrm{EE}}(p)=\sqrt{\frac{N-n}{n-1}\frac{pq}{N}}$$

Como resultado inmediato se tiene el estimador del total de la población con la característica de interés y su varianza.

$$V(\hat{A})=V(Np)=N^2V(p)$$

$$V(\hat{A})=N^2\frac{N-n}{n-1}\frac{PQ}{N}$$

Cuyo estimador es: $$\hat{V}(\hat{A})=\frac{N(N-n)}{n-1}pq$$

2.11 Efecto de P en la varianza del estimador p

Si se ignora el factor de corrección por finitud en la varianza de p, se tiene $V(p)=\frac{PQ}{n}$

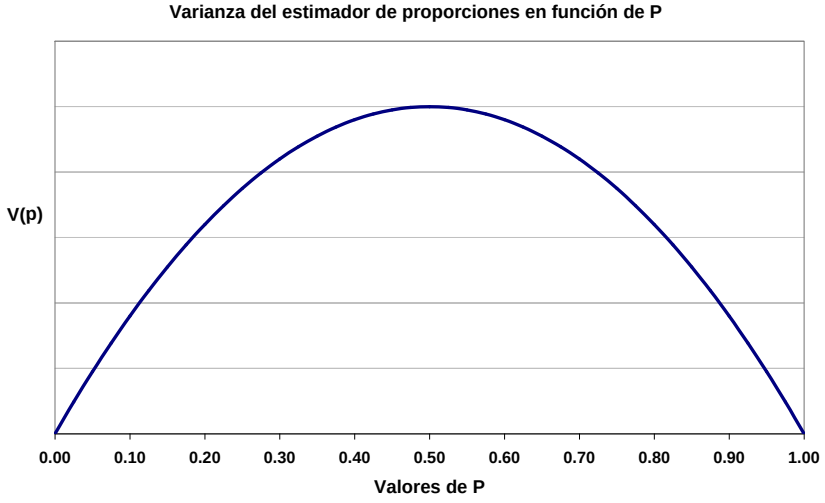
Se toma esa expresión como una función de P

$$\phi(P)=\frac{PQ}{n}=\frac{1}{n}(P-P^2)$$

Se deriva ϕ(P) respecto a P y se iguala a cero para obtener el valor de P que maximiza a ϕ(P)

$$\frac{\partial\phi(P)}{\partial P}=\frac{1-2P}{n}=0 \quad \text{de donde} \quad P=\frac{1}{2}$$

Por lo tanto, V(p) es máxima cuando $P=\frac{1}{2}$



El resultado anterior se suele utilizar para tener una cota máxima de la varianza y en función de ella calcular un tamaño de muestra conservador.

2.12 La distribución Hipergeométrica en relación al estimador p.

Para el caso de p es posible conocer su distribución exacta de al vincularla a una variable aleatoria X que representa el número de elementos de la muestra que pertenecen a la clase de interés. El modelo se enfoca como el caso de una urna con A elementos de la clase C y N-A elementos de la clase complementaria C*. La probabilidad de que al extraer una muestra, sin reemplazo se obtengan X = a elementos de la clase C responde a una distribución hipergeométrica. En este contexto se considera la relación p=a/n de donde X=np.

$$P(X=a)=\frac{\binom{A}{a}\binom{N-A}{n-a}}{\binom{N}{n}} \qquad \begin{array}{l} a=0,1,2,\ldots,n \\ a<A \\ n-a<N-A \end{array}$$

Si se considera A/N = P son inmediatos los siguientes resultados:

$$E(X)=\frac{nA}{N}=nP$$

$$V(X)=\frac{nA(N-A)}{N^2}\frac{(N-n)}{(N-1)}=nP(1-P)\frac{(N-n)}{(N-1)}$$

Como V(X) = V(np) = n^2V(p). La varianza de p se obtiene al dividir V(X) entre n^2 y se llega a la misma fórmula que se obtuvo para V(p) como caso particular de la media de una muestra aleatoria simple.

$$\mathrm{V(p)}=\frac{\mathrm{P(1-P)}}{\mathrm{n}}\frac{(N-n)}{(N-1)}$$

La hipergeométrica se puede aproximar mediante la Binomial cuando A y N-A son grandes en relación al tamaño de muestra n. Mediante la distribución Hipergeométrica y la Binomial se han elaborado tablas y gráficas como las de Chung y De Lury (Confidence limits for the hipergeometric distribution, University of Toronto Press; 1950) para establecer intervalos de confianza para P al 90, 95 y 99% de confianza con N=500, 2500 y 10000. La capacidad de las computadoras personales actuales permite construir estos intervalos para tamaños moderados de los valores parametrales de la población.

2.13 Intervalos de confianza para P mediante la aproximación normal.

Si suponemos que una p se distribuye aproximadamente normal es fácil construir intervalos de confianza basados en esta propiedad. La pregunta fundamental es ¿cuándo se puede suponer la normalidad de p en presencia de determinados valores de N, n, P y el nivel de confianza seleccionado? Se ha verificado que el error de aproximación es más sensible respecto a n y P. La conclusión es que si n es moderadamente grande y P está cercano a 0.5 se puede suponer normalidad para p sin problema, pero para valores alejados de 0.5 la asimetría en la distribución de p juega un papel pernicioso.

W. Cochran da los siguientes valores mínimos de n requeridos para suponer la normalidad de p:

Valores de P		Muestra
Menor a	Mayor a	Requerida
0.40	0.60	50
0.30	0.70	80
0.20	0.80	200
0.10	0.90	600
0.05	0.95	1400

Si se puede suponer la normalidad, los límites de confianza para p se pueden obtener de la expresión siguiente:

$$p \pm \left\{ Z_{(1-\alpha/2)} \sqrt{\frac{N-n}{n-1}\frac{pq}{N}} + \frac{1}{2n} \right\}$$

En esta expresión el cociente $\frac{1}{2n}$ es un factor de corrección por continuidad cuyo efecto es un intervalo más conservador. Sin la aplicación de esta corrección, el intervalo resulta ligeramente más angosto.

Ejemplo 2.3

Supóngase que en una muestra de n =500 estudiantes de una universidad con 20,000 alumnos, 150 de ellos se transportan en auto propio. Construya un intervalo de confianza de 95% para P.

$$p \pm \left(Z_{(1-\alpha/2)} \sqrt{\frac{N-n}{n-1}\frac{pq}{N}} + \frac{1}{2n} \right)$$

$$0.30 \pm \left(1.96 \sqrt{\frac{20{,}000-100}{500-1}\frac{(0.3)(0.7)}{20{,}000}} + \frac{1}{2(500)} \right)$$

$$0.30 \pm (1.96(0.02025637) + 0.001)$$

$$0.30 \pm 0.040702$$

El intervalo requerido es (0.259277, 0.340702)

2.14 Tamaño de Muestra para Proporciones

Si se supone la normalidad se puede obtener una expresión para n, análoga a la obtenida para el caso $\bar{y}$ con una precisión d y confianza 100(1-α)%

$$d = Z_{(1-\alpha/2)} EE_p$$

$$d^2 = Z^2_{(1-\alpha/2)} \frac{N-n}{N-1} \frac{PQ}{n}$$

$$d^2(N-1)n = Z^2_{(1-\alpha/2)}(N-n)PQ$$

$$d^2Nn - d^2n = Z^2_{(1-\alpha/2)}NPQ - Z^2_{(1-\alpha/2)}nPQ$$

$$\left(Z^2_{(1-\alpha/2)}PQ + d^2N - d^2\right)n = Z^2_{(1-\alpha/2)}NPQ$$

$$n = \frac{Z^2_{(1-\alpha/2)}NPQ}{d^2N + Z^2_{(1-\alpha/2)}PQ - d^2}$$

Se divide numerador y denominador entre d^2N:

$$n = \frac{\frac{Z^2_{(1-\alpha/2)}PQ}{d^2}}{1 + \frac{Z^2_{(1-\alpha/2)}PQ}{Nd^2} - \frac{1}{N}}$$

$$n = \frac{\frac{Z^2_{(1-\alpha/2)}PQ}{d^2}}{1 + \left(\frac{Z^2_{(1-\alpha/2)}PQ}{d^2} - 1\right)\frac{1}{N}}$$

Si se identifica $n_o = \frac{Z^2_{(1-\alpha/2)}PQ}{d^2}$ que corresponde al tamaño de una muestra con reemplazo

Finalmente se tiene el tamaño de muestra en función de $\mathbf{n_o}$: $$n = \frac{n_o}{1 + \frac{n_{0-1}}{N}}$$

Ejemplo 2.4

En una muestra preliminar de n = 50 estudiantes seleccionada de una población de N=4000 se encuentra que a = 30 fuman. ¿Qué tan grande debe ser la muestra para estimar p con una precisión de 5% con una confianza de 99%?.

N = 4000

n = 50

$$n = \frac{n_o}{1 + \frac{n_{0-1}}{N}}$$

p=0.6
q=0.4
Z=2.58

d = 0.05 (absoluta)

$$n_o = \frac{Z^{2}_{(1-\alpha/2)} pq}{d^2} = \frac{(2.58)^2 (0.6)(0.4)}{(0.05)^2} = 639.0144$$

$$n = \frac{639.0144}{1 + \frac{639.0144 - 1}{4000}} = 551.11032$$

n = 552

Ejemplo 2.5

Unos antropólogos desean estimar la proporción de personas de una región de 6,000 habitantes que presentan cierta característica de tipo hereditario. No disponen de datos de una prueba piloto y simplemente conjeturan que la característica se presente en la mitad de los habitantes para tener un tamaño conservador de la muestra. Calculan tamaño de muestra para estimar la característica con una precisión de 0.03 y 95% de confianza.

P = q = 0.5

$$n_o = \frac{Z^{2}_{(1-\alpha/2)} pq}{d^2} = \frac{(1.96)^2 (0.5)(0.5)}{(0.03)^2} = 1067.11$$

d = 0.03

Z = 1.96

$$n = \frac{1067.11}{1 + \frac{1067.11 - 1}{6000}} = 906.1$$

Se redondea al entero mayor o igual y por lo tanto **n = 907**