MUESTREO ESTADÍSTICO

SEMANA-4: Tamaños Muestrales en M.A.S

Raúl Alberto Pérez

Universidad Nacional de Colombia, Escuela de Estadística, 2021-I

Tamaños de Muestra en M.A.S

La importancia que tiene el tamaño de muestra ha sido ignorada en muchas situaciones dando lugar a problemas de difícil solución, como por ejemplo, tratar de obtener conclusiones estadísticamente aceptables, cuando la información recolectada es ***insuficiente*** o gran parte de ella es ***redundante*** y no agrega nada a la otra ya registrada.

La falta de una verdadera conciencia acerca del problema del tamaño de la muestra hace que este problema sea abordado cuando ya el estudio está avanzado y no al principio de dicho estudio como debería ser.

Aunque en la práctica son múltiples las variables a considerar y los parámetros a estimar, inicialmente **se asumirá que sólo se tiene una variable y un parámetro de interés** y que el diseño a usar es el M.A.S.

Antes de iniciar el proceso de encontrar el tamaño de la muestra, hay que responder algunas preguntas, como por ejemplo:

Qué tan exacto se desea la estimación del parámetro de interés? precisión

¿Con qué nivel de confiabilidad se desea la estimación? (α)

¿Es la distribución del estimador del parámetro de interés aproximadamente normal?

Se tienen valores de referencia para algunos parámetros poblacionales? σ^2

Es decir, debe de existir algún enunciado con respecto a lo que se espera de la muestra. Por ejemplo límites de error deseados.

Se debe encontrar una ecuación que relacione a n con la precisión deseada de la muestra.

El muestreo probabilístico permite la elaboración de dicha ecuación.

Tamaños de Muestra para la estimación de medias y totales en M.A.S

El procedimiento regular para hallar el tamaño de muestra se basa en considerar una fórmula que incluya dicho tamaño (n) y otros términos conocidos a los que se les puede asignar un valor previo.

La fórmula considerada debe permitir despejar fácilmente a n, aunque en algunos casos se necesitarán métodos numéricos de aproximación.

Conocidos el límite para el error de estimación B y el nivel de confianza $1-\alpha$ fijados por el investigador para estimar μ, se halla cual debe ser el tamaño de muestra n, de la siguiente forma.

Para ello se usa la expresión para el Límite del error de estimación dada por:

$$B = Z_{\alpha/2}\sqrt{Var[\overline{y}]} = Z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)},$$

de donde despejando a n, se obtiene que:

$$n = \frac{\sigma^2 N}{\frac{B^2}{Z^2_{\alpha/2}}(N-1)+\sigma^2} \tag{1}$$

Haciendo:

$$D = \frac{B^2}{Z^2_{\alpha/2}}$$

se tiene al expresión alternativa:

$$n = \frac{\sigma^2 N}{D(N-1)+\sigma^2} \tag{2}$$

Note que la expresión anterior dada en la ecuación (1) se puede reescribir como:

$$n = \frac{1}{\frac{1}{N} + \left(\frac{N-1}{N}\right)\frac{1}{n_0}}, \quad \text{con:} \quad n_0 = \frac{Z^2_{\alpha/2}\sigma_0^2}{B^2_\mu} \tag{3}$$

Esta última expresión depende de:

- N: Número de unidades en el MARCO.
- σ^2: Varianza poblaciónal de la variable de interés. Si no se conoce se estima usando: $\sigma_0^2 \approx S^2$.
- B: Límite en el error de Estimación para la media poblacional.
- $1-\alpha$: Nivel de confianza: $P\Big[|\hat{\mu} - \mu| \leq B\Big] = 1-\alpha$.
- n_0: Tamaño de muestra cuando la población es INFINITA.

Tamaño de Muestra cuando se desea controlar el Error Máximo Relativo (EMR)

En muchos casos es más conveniente tratar de controlar el ***Error Máximo Relativo*** *(EMR)*, en lugar del EMA.

Estos casos se dan cuando hay un desconocimiento de alto grado de la distribución poblacional de la variable de interés y el resultado de la muestra se vuelve incierto.

Por ejemplo, en un estudios de ingresos un EMA B de $\$100000$, puede ser grande o pequeño, dependiendo de qué tan simétrica es la distribución de la población, de qué tan grande son los ingresos de las personas, y también, de qué tan válida es la aplicación del TLC.

Si todas estas situaciones anteriores son desconocidas, entonces se podría pecar por exceso o por defecto y esto podrían afectar sensiblemente los resultados finales.

En el mismo ejemplo de los ingresos, el establecimiento de un Error Relativo del $10\,\%$, con respecto al verdadero valor de la media poblacional, podría ser una solución alternativa más aceptable, pues, independientemente de cual sea la distribución original de la población y asumiendo que la aproximación normal es aceptable, siempre se garantizará con algún nivel de confiabilidad, que la estimación no será ni mayor ni menor que la verdadera media poblacional en más de un $10\,\%$, independientemente de si los valores muestrales obtenidos son mayores o menores que los esperados.

El principal problema radica en que cuando se utiliza el error relativo, los tamaños muestrales estimados son generalmente mayores que cuando se utiliza el error absoluto.

Tamaño de muestra para estimar a μ con otras consideraciones

1. El Error Máximo Relativo para estimar μ, está dado por:

$$\epsilon = \frac{B_\mu}{\mu}, \qquad \text{es decir que:} \quad B_\mu = \epsilon\mu$$

2. Relación entre el coeficiente de variación y error máximo relativo:

$$CV = \frac{\sigma}{\mu}, \qquad \text{se tiene que :} \quad B_\mu = \epsilon\frac{\sigma}{CV}$$

Estas expresiones son reemplazadas en las fórmulas anteriores para obtener los respectivos tamaños de muestra.

La ventaja que tiene la fórmula que utiliza el coeficiente de variación, es que en algunas situaciones puede ser más sencillo hacer una estimación previa de la homogeneidad de la población estudiada, que una estimación de su varianza.

Lo anterior se puede lograr mediante información previa acerca de la población en estudio, o a partir de poblaciones similares, las cuales permiten hacer una estimación razonable del coeficiente de variación poblacional, ya que la homogeneidad de una población no cambia tan rápidamente como lo hace la varianza.

Tamaño de muestra para estimar el total poblacional τ

Note que la expresión dada en la ecuación (3) se puede reescribir como sigue, para el caso de la estimación de τ, con Error Máximo Absoluto: B_τ:

$$n = \frac{1}{\frac{1}{N} + \left(\frac{N-1}{N}\right)\frac{1}{n_0}}, \quad \text{con:} \quad n_0 = \frac{N^2 Z^2_{\alpha/2}\sigma_0^2}{B_\tau^2} \tag{4}$$

La expresión anterior depende de:

- N: Número de unidades en el MARCO.
- σ^2: Varianza poblacional de la variable de interés. Si no se conoce se estima usando: $\sigma_0^2 \approx S^2$.
- B_τ: Límite en el error de Estimación para el total poblacional.
- $1-\alpha$: Nivel de confianza $P\Big[|\widehat{\tau} - \tau| \leq B\Big] = 1-\alpha$.
- n_0: Tamaño de muestra cuando la población es INFINITA.

Tamaño de muestra para la estimación de τ con otras consideraciones

1. Error Máximo Relativo para estimar τ:

$$\epsilon = \frac{B_\tau}{\tau}, \quad \text{de donde:} \quad B_\tau = \epsilon\tau = \epsilon N\mu.$$

2. Relación entre el coeficiente de variación y Error Máximo Relativo:

$$CV = \frac{\sigma}{\mu}, \quad \text{de donde:} \quad B_\tau = N\epsilon\frac{\sigma}{CV}.$$

Tamaño de muestra para la estimación de la proporción poblacional p

Como ya se ha mencionado, la proporción es un caso particular de la media, por consiguiente todo el desarrollo llevado a cabo para hallar tamaños de muestra para la media es igualmente válido para proporciones.

Debe tenerse en cuenta, que la proporción muestral debe seguir aproximadamente una distribución normal, y para que esto ocurra, deben satisfacerse los valores mínimos de n dados en la tabla:

p	0.5	0.4	0.3	0.2	0.1	0.05	<0.05
n	30	50	80	200	600	1400	>1400

Cuando esto no se cumple, se debe utilizar la distribución Hipergeométrica o la Binomial para hallar los tamaños de muestra.

Se supondrá que esta condición de normalidad se cumple.

Tamaño de muestra para la estimación de la proporción poblacional p

El tamaño de muestra n para la estimación de p con un Límite para el Error Máximo Absoluto de B_p, se halla a partir de (se reemplaza σ^2 por $p(1-p)$):

$$n = \frac{1}{\frac{1}{N} + \left(\frac{N-1}{N}\right)\frac{1}{n_0}}, \quad \text{con:} \quad n_0 = \frac{Z^2_{\alpha/2} p(1-p)}{B_p^2}$$

La expresión anterior depende de:

- N: Número de unidades en el MARCO.
- p: Proporción Poblacional, se estima a partir de un estudio piloto.
- B_p: Límite en el error de Estimación para la proporción poblacional.
- $1-\alpha$: Nivel de confianza $P\Big[|\widehat{p} - p| \leq B\Big] = 1-\alpha$.
- n_0: Tamaño de muestra cuando la población es INFINITA.

NOTA: El tamaño de muestra más grande se alcanza cuando $p = 0{,}5$.

EJEMPLO:

Tamaño de muestra CASO DE ESTUDIO: Estimación de μ.

Hallar n con el fin de estimar el Valor de la matrícula promedio con un Límite en el error de Estimación de $B = 200000$ y un nivel de confianza del $95\,\%$, use como estimación para σ_0^2 el valor de S^2 hallado en la muestra inicial dado por: $\sigma_0^2 \approx 434293,89 \times 1000^2$.

En este caso:

$$n_0 = \frac{Z^2_{\alpha/2}\sigma^2}{B^2} = \frac{Z^2_{0,05/2}434293,89 \times 1000^2}{200000^2} = \frac{1{,}96^2 \times 434293,89 \times 1000^2}{200000^2} = 41,7095$$

luego se tiene que:

$$n = \frac{1}{\frac{1}{N} + \left(\frac{N-1}{N}\right)\frac{1}{n_0}} = \frac{1}{\frac{1}{739} + \left(\frac{739-1}{739}\right)\frac{1}{41{,}7095}} = 39{,}5318 \approx 40$$

Tamaño de muestra CASO DE ESTUDIO: Estimación de la Proporción: p.

En este caso se reemplaza a σ^2 por $p(1-p)$.

Calcular el número de estudiantes a muestrear para estimar la Proporción de estudiantes del primer semestre de la Facultad de Minas que llegan a la Unal en bicicleta con un Límite en el Error de estimación de $B = 5\,\%$, usar $p = 0{,}28$, valor estimado con la muestra inicial.

Ene ste caso se tiene que:

$$n = \frac{p(1-p)N}{\frac{B^2}{Z^2_{\alpha/2}}(N-1) + p(1-p)} = \frac{0{,}28(1-0{,}28)739}{\frac{0{,}05^2}{1{,}96^2}(739-1) + 0{,}28(1-0{,}28)} = 218{,}49 \approx 219$$

o de forma equivalente se puede usar:

$$n = \frac{1}{\frac{1}{N} + \left(\frac{N-1}{N}\right)\frac{1}{n_0}}, \quad \text{con:} \quad n_0 = \frac{Z^2_{\alpha/2}p(1-p)}{B_p^2}$$

Importancia del efecto de diseño en la determinación del tamaño muestral

El diseño M.A.S es tomado como referencia cuando se desean utilizar otros métodos de estimación correspondientes a diseños más complejos.

La eficiencia de un diseño específico, que se denotará por, EFD, es la eficiencia relativa que el estimador, usando este diseño, tiene con respecto al estimador del M.A.S, en la estimación de un parámetro, en generla es:

$$EFD = \frac{Var[\widehat{\theta}]}{Var[\widehat{\theta}_{MAS}]} \tag{5}$$

La expresión anterior, se puede usar para estimar el tamaño de muestra necesario de acuerdo con el nuevo diseño, para lograr la misma eficiencia que con el M.A.S.

Después de un poco de álgebra se obtiene que:

$$n = \frac{(EFD)\left(\frac{z^2\sigma^2}{B_\mu^2}\right)}{1 + (EFD)\left(\frac{z^2\sigma^2}{NB_\mu^2}\right)}. \tag{6}$$

como en el caso del M.A.S, se debe estimar la varianza poblacional σ^2.

Tamaño Muestra Caso Estudio: Efecto de Diseño

Usando $p = 0{,}28$, $S^2 = pq = 0{,}28 * 0{,}72$, $B_p = 0{,}05$, $z_{0{,}05/2} = 1{,}96$, $N = 739$.

Con diferentes valores de Efecto de Diseño, **EFD**:

$$n = \frac{(\mathbf{EFD})\left(\frac{z^2 * p * (1-p)}{B_p^2}\right)}{1 + (\mathbf{EFD})\left(\frac{z^2 * p * (1-p)}{N B_p^2}\right)} = \frac{(\mathbf{EFD})\left(\frac{1{,}96^2 * 0{,}28 * (0{,}72)}{0{,}05^2}\right)}{1 + (\mathbf{EFD})\left(\frac{1{,}96^2 * 0{,}28 * (0{,}72)}{739 * 0{,}05^2}\right)}$$

$$n = \frac{(\mathbf{EFD}) * 309{,}786624}{1 + (\mathbf{EFD})\frac{309{,}786}{739}} = \frac{\mathbf{EFD}}{\frac{1}{309{,}786} + \frac{\mathbf{EFD}}{739}}$$

EFD	0.9	1.0	1.1	1.2	1.3	1.4	1.5
n	203	**219**	234	248	261	274	286

Procedimientos sugeridos para la estimación previa de la varianza poblacional σ^2

EL principal problema presente cuando se desea estimar un tamaño muestral controlando el EMA B, es el conocimiento aproximado de la varianza poblacional σ^2.

Existen varios caminos para estimar esta varianza poblacional, entre los cuales están los siguientes.

1. Revisión bibliográfica de estudios anteriores sobre la misma población o poblaciones similares. A juicio del investigador se decidirá cual de las estimaciones disponibles es la más apropiada, teniendo en cuenta factores como el tiempo transcurrido, el tipo de variable analizada, si la población es o no es la misma, etc.

2. Selección de una muestra piloto de tamaño n_1. Generalmente menor a 30. E estima la varianza poblacional con la varianza de esa muestra, la cual se representa por S^2 en la respectiva fórmula.

   Si esta muestra piloto no se selecciona de una forma completamente aleatoria, entonces los elementos que forman parte de ella NO deben ser parte de la muestra definitiva.

   En algunos casos, la muestra aleatoria se puede seleccionar de manera intencional atendiendo criterios de expertos en el tema.

3. **Selección de una m.a.s** de tamaño n_1 y cálculo de la varianza de esta muestra. Esta varianza se toma como estimación de la varianza poblacional.

   Sin embargo, el hecho de estimar la varianza a partir de una muestra aleatoria pequeña conlleva a un margen de incertidumbre que se supone puede corregirse multiplicando el valor obtenido de n por un factor que depende de n_1.

La fórmula final para n está dada por:

$$n = \left(\frac{z^2 S^2}{B_\mu^2}\right)\left(1 + \frac{2}{n_1}\right).$$

La cantidad $\left(\frac{2}{n_1}\right)$-es el precio que se paga por el desconocimiento de σ^2.

NOTA: Para el caso de la estimación de una proporción, $S^2 = p(1-p)$.

En el caso de estudio, el tamaño de muestra, usando los $n_1 = 25$ estudiantes, estará dado por:

$$n = \left(\frac{1{,}96^2 0{,}28 * 0{,}72}{0{,}05^2}\right)\left(1 + \frac{2}{25}\right) = 322,17 \approx 323$$

Por lo tanto, se requieren $323 - 25 = 298$ nuevas encuestas para alcanzar un límite en el error de estimación del $5\,\%$, para la estimar la proporción de estudiantes de recién ingreso que utilizan la bicicleta para la llegar a la Universidad.

Una desventaja tanto de este procedimiento como del anterior, es la demora que se puede presentar en la recolección de toda la información, sin embargo es el procedimiento más usado entre los propuestos.

4. Una alternativa similar a la del numeral (3) es presentada por Desu y Raghavarao (1990).

   En este caso se hace uso de la distribución t de Student, en la expresión para hallar n:

$$n = \text{máx}\left\{ n_1 \ , \ \left[\frac{t^2_{\alpha/2;n_1-1} S^2}{B^2_\mu} \right] \right\} .$$

Para el caso de estudio se tiene que:

$$n = \text{máx}\left\{ n_1 \, , \, \left[\frac{t^2_{\alpha/2;n_1-1} S^2}{B^2_\mu} \right] \right\}$$

$$= \text{máx}\left\{ 25 \, , \, \left[\frac{t^2_{0,05/2;25-1} 0{,}28 * (1 - 0{,}28)}{0{,}05^2} \right] \right\}$$

$$= \text{máx}\left\{ 25 \, , \, \left[\frac{2{,}063899^2 * 0{,}28 * (1 - 0{,}28)}{0{,}05^2} \right] \right\}$$

$$= \text{máx}\{25 \, , 344\}$$

$$n = 344.$$

5. Determinación tentativa, o con base en supuestos adecuados, de la estructura de la población para escoger la distribución teórica que mejor podría representarla (normal, exponencial, uniforme, etc.).

   Al identificar una distribución apropiada se usan sus propiedades para obtener una estimación más realista de la varianza.

   Si el desconocimiento es total se debe recurrir a la distribución uniforme.

EJEMPLO: Considere la información que aparece en la siguiente tabla, la cual corresponde a una muestra de 30 personas. Una vez las personas fueron seleccionadas, además de su ingreso mensual (I), se registró su género (G) (masculino m, femenino f) y su estado civil (EC) (soltero s, casado c, otro (o).

Nro.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	f	m	f	m	f	m	f	f	f	m	f	f	m	m	m
EC	c	s	c	s	c	c	c	o	o	s	o	s	s	s	c
I	2	2.5	4	3.8	7.2	10	5.6	4.9	3.3	4	3.5	5.7	10	8.1	4.4

Nro.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
G	f	m	f	f	m	m	f	m	f	m	f	m	f	m	f
EC	o	s	o	c	s	c	s	o	c	s	o	s	o	c	s
I	6.6	7.3	8	9	3.9	7.1	4.9	2.3	3.9	11.1	7.3	6.5	5.8	4	3

Resolver las siguientes preguntas:

1. Supóngase que la muestra es una muestra piloto seleccionada por expertos de una comunidad de 10000 personas. Se desea determinar el tamaño de muestra mínimo para estimar el ingreso promedio (y total) de la comunidad con un error absoluto no mayor de 0.2 salarios mínimos (y de 2000 salarios para el caso del total) y una confiabilidad mínima del $95\,\%$.

   Estamos ante una situación donde se debe usar la siguiente expresión:

$$n = \frac{NS^2}{(N-1)D + S^2}, \quad \text{con: } D = \frac{B^2}{Z^2_{\alpha/2}}$$

   Primero identifiquemos cada término de dicha expresión:

   Para una confiabilidad del $95\,\%$ se tiene que: $z = 1{,}96$.

   Para un EMA no mayor a $0{,}2$ se tiene que: $B = 0{,}2$ salarios mínimos.

   Para $B = 0{,}2$ se tiene que: $D = \frac{B^2}{z^2} = \frac{0{,}2^2}{1{,}96^2} = 0{,}0104$.

   $S^2 = 6{,}0590$, lo cual es la varianza de la muestra piloto de 30.

   El tamaño de la población es: $N = 10000$-personas.

Reemplazando todos estos valores se tiene que:

$$n = \frac{NS^2}{(N-1)D + S^2} = \frac{10000(6{,}0590)}{9999(0{,}0104) + 6{,}0590} = 549{,}91 \approx 550$$

es decir, se debe seleccionar a 550 personas para que las estimaciones cumplan los requisitos exigidos y además no se deben incluir en la muestra final las personas de la muestra piloto.

2. Si lo que se desea controlar es el error máximo relativo (ε), en lugar del EMA (B), y se busca que dicho EMR sea como máximo del $5\,\%$ y adicionalmente se asume que se conoce una estimación a priori de la media y varianza, que están dadas por las de la muestra piloto, hallar el tamaño n necesario.

En este caso se debe usar la expresión:

$$n = \frac{NS^2}{(N-1)D + S^2}$$

$$\text{con,} \quad D = \frac{B^2}{z^2}, \quad \text{y}$$

$$B^2 = \mu^2\varepsilon^2 = \overline{y}^2\varepsilon^2 = (5{,}66)^2(0{,}05)^2 = 0{,}080089, \quad \text{de donde:}$$

$$D = \frac{B^2}{z^2} = \frac{0{,}080089}{(1{,}96)^2} = 0{,}02084$$

y por lo tanto:

$$n = \frac{NS^2}{(N-1)D + S^2} = \frac{10000(6{,}0590)}{9999(0{,}02084) + 6{,}0590} = 282{,}05 \approx 283$$

3. Asumamos que lo único que se conoce es que ninguna persona de la comunidad tiene ingreso superior a 15 salarios mínimos y que la distribución de ingresos de la comunidad es aproximadamente normal.

   Para estimar el tamaño de muestra en este caso, con la misma precisión absoluta (B) y la misma confiabilidad (z) de los puntos anteriores, es necesario considerar algunas características de la distribución normal.

   La característica más importante en este caso, es que entre la media y más o menos tres desviaciones estándar se encuentran el $99{,}73\,\%$ de los valores. Para fines prácticos este se considera igual al $100\,\%$.

   Ahora, la diferencia entre los dos valores extremos del intervalo es aproximadamente igual al rango de la variable, que en el ejemplo sería 15.

Igualando este valor al rango, que a su vez corresponde a 6-desviaciones estándar, se puede encontrar una estimación previa de la desviación estándar poblacional (y por lo tanto de la varianza poblacional), como sigue:

$$6S = 15\ , \quad \text{de donde} \quad S = 2{,}50,$$

y por lo tanto: $S^2 = 2{,}5^2 = 6{,}25$.

Reemplazando se obtiene:

$$n = \frac{NS^2}{(N-1)D + S^2} = \frac{10000(6{,}25)}{9999(0{,}0104) + 6{,}25} \approx 567$$

4. Suponga que no se puede decir nada de la distribución de los ingresos.

   En este caso la alternativa más indicada es asumir que la variable de estudio (el ingreso) se distribuye uniformemente y usar las propiedades de esta distribución para encontrar una estimación de la varianza.

   Si el mayor valor que toma la variable es 15-salarios mínimos y el menor es de 0, estos dos valores corresponden a los extremos de la distribución.

   La varianza de la distribución uniforme, $U(a, b)$, en nuestro caso, $U(0, 15)$, está dada por:

$$S^2 = Var[U] = \frac{(b-a)^2}{12} = \frac{(15-0)^2}{12} = 18{,}75.$$

   Luego reemplazando se tiene que:

$$n = \frac{NS^2}{(N-1)D + S^2} = \frac{10000(18{,}75)}{9999(0{,}0104) + 18{,}75} \approx 1526$$

EJEMPLO: La información de la siguiente tabla, corresponde a una muestra piloto de 30 personas de una población de $N = 10000$ personas. Una vez las personas fueron seleccionadas, además de su ingreso mensual (I), se registró su género (G) (masculino m, femenino f) y su estado civil (EC) (soltero s, casado c, otro (o).

Nro.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	f	m	f	m	f	m	f	f	f	m	f	f	m	m	m
EC	c	s	c	s	c	c	c	o	o	s	o	s	s	s	c
I	2	2.5	4	3.8	7.2	10	5.6	4.9	3.3	4	3.5	5.7	10	8.1	4.4

Nro.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
G	f	m	f	f	m	m	f	m	f	m	f	m	f	m	f
EC	o	s	o	c	s	c	s	o	c	s	o	s	o	c	s
I	6.6	7.3	8	9	3.9	7.1	4.9	2.3	3.9	11.1	7.3	6.5	5.8	4	3

Se desea estimar el porcentaje (proporción) de hombres, con un error absoluto no mayor de 0.04 y una confiabilidad de aproximadamente el $95\,\%$.

Usando la información de la muestra piloto, se puede hacer una estimación preliminar de P, la cual se denota por P_0 y está dada por:

$$P_0 = \frac{a}{n} = \frac{14}{30} = 0{,}47, \qquad B = \frac{B^2}{4} = \frac{(0{,}04)^2}{4} = 0{,}0004$$

Luego se tiene que:

$$n = \frac{NP_0Q_0}{(N-1)D + P_0Q_0} = \frac{10000(0{,}47)(0{,}53)}{(9999)(0{,}0004) + (0{,}47)(0{,}53)} = 586{,}3$$

es decir que el número mínimo de personas a seleccionar es de: $n = 587$.

Si la muestra inicial fue aleatoria, ésta puede hacer parte de la muestra final.

Además, $n_0 = 625$

Tamaño de Muestra para estimar varias Proporciones Simultáneamente

Suponga que se tiene una población finita, pero grande, de tamaño N, de la cual se conoce que está dividida en k-categorías mutuamente excluyentes (caso típico de las encuestas de opinión), pero se desconoce la participación de cada una de dichas categorías en la población, ie. las $P_j, j = 1, 2, \cdots, k$, son desconocidas.

La aproximación multinomial ha sido la herramienta más utilizada en la solución de este problema.

Existen varios métodos que analizan tamaños muestrales en poblaciones multinomiales, los cuales asumen que N-es suficientemente grande para ignorar el factor de corrección por población finita (cpf) si el muestreo se hace sin reemplazo, además suponen que los tamaños de las muestras también son suficientemente grandes.

Thompson (1987) [*] demostró que la peor situación ocurre cuando m de las k-proporciones ($m \leq k$) son iguales entre sí y el resto de ellas son cero, ie. $P_1 = P_2 = \cdots = P_m$, $P_{m+1} = P_{m+2} = \cdots = P_{k-m} = 0$.

El valor de m depende del nivel de significancia α.

A continuación se presentan los valores de n_0 sugeridos, correspondientes a diferentes combinaciones del nivel de significancia α y del EMA $B = 0{,}05$:

[*] Thompson, S. K. (1987). Sample size for estimating multinomial proportions. *The American Statistician*, **41**, 42-46.

α	Máx-$B^2 n_0 = (0{,}05)^2 n_0$	n_0	m
0.5	0.44129	177	4
0.4	0.50729	203	4
0.3	0.60123	241	3
0.2	0.74739	299	3
0.10	1.00635	403	3
0.05	1.27359	510	3
0.025	1.55963	624	2
0.020	1.65872	664	2
0.010	1.96986	788	2
0.005	2.28514	915	2
0.001	3.02892	1212	2

El valor de n_0 en la tabla anterior se determina a partir de:

$$n_0 B^2 = \max_m \left\{ \left[\Phi^{-1}\left(1 - \frac{\alpha}{2m}\right) \right]^2 \frac{1}{m}\left(1 - \frac{1}{m}\right) \right\}$$

una vez identificada el valor de n_0 se ajusta por la finitud del tamaño de la población:

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

(ver código de R).

EJEMPLO: Considere (nuevamente) la información que aparece en la siguiente tabla, la cual corresponde a una muestra de 30 personas. Una vez las personas fueron seleccionadas, además de su ingreso mensual (I), se registró su género (G) (masculino m, femenino f) y su estado civil (EC) (soltero s, casado c, otro (o).

Nro.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	f	m	f	m	f	m	f	f	f	m	f	f	m	m	m
EC	c	s	c	s	c	c	c	o	o	s	o	s	s	s	c
I	2	2.5	4	3.8	7.2	10	5.6	4.9	3.3	4	3.5	5.7	10	8.1	4.4

Nro.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
G	f	m	f	f	m	m	f	m	f	m	f	m	f	m	f
EC	o	s	o	c	s	c	s	o	c	s	o	s	o	c	s
I	6.6	7.3	8	9	3.9	7.1	4.9	2.3	3.9	11.1	7.3	6.5	5.8	4	3

Se desea estimar el porcentaje (o proporción) de personas solteras, casadas y con otro estado civil, con un error absoluto máximo del $B = 0{,}05$ para cada una de las categorías y un nivel de confiabilidad del $90\,\%$.

Usando la tabla de Thompson, con un $\alpha = 0{,}1$ y $m = 3$, se encuentra que el valor de $n_0 B^2 = 1{,}00635$, de donde despejando se tiene que: $n_0 = 403$.

Al ajustar este valor por la finitud de la población se tiene que:

$$n = \frac{n_0}{\frac{N-1}{N} + \frac{n_0}{N}} = \frac{403}{\frac{9999}{10000} + \frac{403}{10000}} = 387{,}43 \approx 388.$$

Tamaños de Muestra en Subpoblaciones

Existen dos alternativas a considerar:

(a) Tamaños de Muestra cuando se desean estimaciones en subpoblaciones No-previamente identificadas.

Al seleccionar una m.a.s sin reemplazo de tamaño n, el número esperado de elementos que pertenecerán a la sub-población k-ésima, $k = 1, 2, \cdots, L$, está dado por:

$$E[n_k] = nP_k = n\left(\frac{N_k}{N}\right),$$

con P_k-la verdadera proporción de elementos de la población que pertenecen a la k-ésima sub-población.

La varianza aproximada del estimador de la media de la k-ésima sub-población está dada por:

$$Var[\overline{y}_k] \approx \frac{\sigma_k^2}{nP_k}$$

El tamaño de muestra mínimo que garantiza una variabilidad dada, para las estimaciones de las medias en cada una de las sub-poblaciones es:

$$n = \underset{k}{\text{Máx}} \left\{ \frac{S_k^2}{Var[\overline{y}_k]P_k} \right\}$$

donde, $Var[\overline{y}_k]$-es una cantidad propuesta inicialmente por el investigador y S_k^2-es la estimación de σ_k^2.

Si las P_k-se desconocen, deberán estimarse a partir de una muestra piloto o a partir de información secundaria.

Ahora, si en lugar de estimar las medias de las subpoblaciones, se desea estimar las diferencias de medias de dichas subpoblaciones, entonces la varianza entre la diferencia de medias es:

$$Var\left[\overline{y}_k - \overline{y}_m\right] = \frac{\sigma_k^2}{n_k} + \frac{\sigma_m^2}{n_m}$$

y el tamaño de muestra aproximado es:

$$n = \underset{k,m}{\text{Máx}} \left\{ \left(\frac{1}{Var\left[\overline{y}_k - \overline{y}_m\right]} \right) \left(\frac{S_k^2}{P_k} + \frac{S_m^2}{P_m} \right) \right\}$$

EJEMPLO: Considere (nuevamente) la información que aparece en la siguiente tabla, la cual corresponde a una muestra de 30 personas. Una vez las personas fueron seleccionadas, además de su ingreso mensual (I), se registró su género (G) (masculino m, femenino f) y su estado civil (EC) (soltero s, casado c, otro (o).

Nro.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	f	m	f	m	f	m	f	f	f	m	f	f	m	m	m
EC	c	s	c	s	c	c	c	o	o	s	o	s	s	s	c
I	2	2.5	4	3.8	7.2	10	5.6	4.9	3.3	4	3.5	5.7	10	8.1	4.4

Nro.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
G	f	m	f	f	m	m	f	m	f	m	f	m	f	m	f
EC	o	s	o	c	s	c	s	o	c	s	o	s	o	c	s
I	6.6	7.3	8	9	3.9	7.1	4.9	2.3	3.9	11.1	7.3	6.5	5.8	4	3

Se desea estimar un tamaño de muestra global que garantice la estimación del ingreso promedio tanto de mujeres (subpoblación-1) como hombres (subpoblación-2), con una variabilidad para las medias muestrales no mayor de 0.01 (lo que equivale a un error absoluto de 0.2 salarios mínimos y a una confiabilidad del 95,45 %).

Supóngase además, que la proporción de hombres y mujeres en la población es igual (ie. $P_1 = P_2 = 0{,}5$).

De la información de los datos muestrales se tiene que:

$$S_1^2 = 5{,}1593 \quad \textbf{y} \quad S_2^2 = 6{,}7979.$$

De los supuestos del problema se tiene que:

$$Var[\overline{y}_1] = Var[\overline{y}_2] = 0{,}01, \quad (= B^2/Z_{\alpha/2}^2).$$

luego: $$n = \underset{k}{\text{Máx}} \left\{ \frac{S_k^2}{Var[\overline{y}_k] P_k} \right\}$$

$$= \underset{1,2}{\text{Máx}} \left\{ \frac{5{,}1593}{(0{,}01)(0{,}5)} \, , \, \frac{6{,}7979}{(0{,}01)(0{,}5)} \right\}$$

$$= \underset{1,2}{\text{Máx}} \{1032 \, , \, 1360\} = 1360.$$

Ahora, si lo que se desea es comparar la diferencia entre los dos ingresos promedios, con una variabilidad prefijada en 0.02 para la diferencia de las estimaciones de los promedios, entonces se tiene que:

$$n = \frac{1}{0{,}02} \left(\frac{5{,}1593}{0{,}5} + \frac{6{,}7979}{0{,}5} \right) = 1195{,}72 \approx 1196.$$

Donde sólo se está considerando una diferencia (entre las subpoblaciones 1 y 2) y por lo tanto no se habla de máximo.

Para la estimación de totales de variables, los tamaños de muestra son los mismos anteriores, siempre que se estén considerando subpoblaciones suficientemente grandes.

En el caso de proporciones se hace de manera similar, teniendo en cuenta las respectivas estimaciones de σ_k^2, utilizando P_k y Q_k.

(b) Tamaños de Muestra cuando se desean estimaciones en subpoblaciones previamente identificadas.

Si las subpoblaciones son identificadas previamente, se debe determinar para cada una de ellas, el tamaño de muestra necesario que permita obtener las estimaciones de los parámetros con los niveles de precisión y confiabilidad deseados.

En este caso se tiene la siguiente expresión de n_k para cada una de las subpoblaciones:

$$n_k = \frac{N_k S_k^2}{(N_k - 1)D + S_k^2} = \frac{1}{\frac{1}{N_k} + \left(\frac{N_k - 1}{N_k}\right)\frac{1}{n_{k0}}}$$, con: $n_{k0} = \frac{Z_{k,\alpha/2}^2 S_{k0}^2}{B_k^2}$

$k = 1, 2, \cdots, L$.

Cada uno de los términos que componen la expresión anterior tienen el mismo significado visto anteriormente en el caso general, pero ahora están limitados a la subpoblación k-ésima.

El tamaño de muestra global n, será la suma de los tamaños de muestras parciales n_k, es decir:

$$n = \sum_{k=1}^{L} n_k = n_1 + n_2 + \cdots + n_L$$

Esta última expresión se puede simplificar en algunos casos, por ejemplo, si se desea la misma precisión y la misma confiabilidad para todas las medias de las k-subpoblaciones, ie. $B_k^2 = B^2$ y $z_k^2 = z^2$.

Si adicionalmente, las subpoblaciones son lo suficientemente grandes $N_k >>>$ como para ignorar el factor cpf y su variabilidad es aproximadamente igual ie. $S_k^2 = S^2$ en todas la k-subpoblaciones, entonces el tamaño global n se reduce a:

$$n = L\left(\frac{z^2 S^2}{B^2}\right) = L(n_k)$$

donde, S^2-es una estimación aproximada de la variabilidad promedio en las subpoblaciones.

EJEMPLO: Considere (nuevamente) la información que aparece en la siguiente tabla, la cual corresponde a una muestra de 30 personas. Una vez las personas fueron seleccionadas, además de su ingreso mensual (I), se registró su género (G) (masculino m, femenino f) y su estado civil (EC) (soltero s, casado c, otro (o).

Nro.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
G	f	m	f	m	f	m	f	f	f	m	f	f	m	m	m
EC	c	s	c	s	c	c	c	o	o	s	o	s	s	s	c
I	2	2.5	4	3.8	7.2	10	5.6	4.9	3.3	4	3.5	5.7	10	8.1	4.4

Nro.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
G	f	m	f	f	m	m	f	m	f	m	f	m	f	m	f
EC	o	s	o	c	s	c	s	o	c	s	o	s	o	c	s
I	6.6	7.3	8	9	3.9	7.1	4.9	2.3	3.9	11.1	7.3	6.5	5.8	4	3

Asuma que la información suministrada en la anterior tabla, corresponde a dos muestras pilotos, una de la subpoblación de mujeres y la otra de las subpoblación de hombres.

Se desea determinar el tamaño de muestra que permita estimar el ingreso promedio de los hombres y de las mujeres separadamente con un error absoluto no mayor de $0{,}2$-salarios mínimos y una confiabilidad del $95\,\%$.

Se considera además que las subpoblaciones son muy grandes y que la variabilidad dentro de ellas es aproximadamente la misma.

En este caso se tiene que:

$$z = 1{,}96\,, \quad B = 0{,}2\,, \quad L = 2$$

EL valor para S^2-se puede tomar como el promedio de las dos estimaciones de varianzas para cada una de las subpoblaciones, ie.

$$S^2 = \frac{S_1^2 + S_2^2}{2} = \frac{5{,}1593 + 6{,}7979}{2} = 5{,}9786$$

Luego,

$$n_1 = n_2 = \frac{z^2 S^2}{B^2} = \frac{(1{,}96)^2(5{,}9786)}{(0{,}2)^2} = 574{,}18 \approx 575$$

y

$$n = L\left(\frac{z^2 S^2}{B^2}\right) = 2(575) = 1150$$