Tema 2. Muestreo Aleatorio Simple

Contenido

1) Definición, selección y notación
2) Estimadores de la media y el total. Propiedades. Varianzas y error de estimación. Límites de confianza.
3) Muestreo simple con restitución
4) Estimadores de la proporción y el total. Varianzas y formas de distribución. Límites de confianza.
5) Tamaño de la muestra para la proporción, la media y el total, con una y varias características.
6) Estimaciones sobre dominios, medias, totales y proporciones.

Muestreo Aleatorio Simple: Definición

El muestreo aleatorio simple, muestreo irrestricto aleatorio o muestreo aleatorio sin restitución, en poblaciones finitas, consiste en seleccionar una muestra de **n** elementos entre los **N** que forman la población, de modo que cada una de las muestras distintas $\left(C_{N,n} = \binom{N}{n} \right)$ tenga la misma probabilidad $1 \Big/ \binom{N}{n} = \frac{n!(N-n)!}{N!}$ de ser seleccionada.

La garantía de equiprobabilidad de la muestra la da la forma aleatoria como se selecciona cada elemento de la muestra. El muestreo aleatorio simple puede ser también con restitución.

Probar que la equiprobabilidad de selección de un elemento específico en la muestra es n/N en el m.a.s. con y sin restitución.

¿Cómo seleccionar la muestra?

1. Tabla de números aleatorios
   ¿Qué es una tabla de números aleatorios?
   ¿Cómo se construye?
   ¿Cómo se conoce su bondad?
   ¿Cómo se emplean?
2. Números seudo-aleatorios
   Generadores de números aleatorios
   Posibilidades de los paquetes estadísticos

Ejemplos: elegir muestras aleatorias de los:
a. alumnos de la Escuela de Estadística
b. suscriptores de teléfono en Ejido
c. pacientes del Hospital Universitario
d. usuarios del mercado principal

Notación:

N tamaño de la población
n tamaño de la muestra
f=n/N fracción de muestreo
N/n factor de expansión (elevación o inflación)

	Poblacional	muestral
Media	$\bar{Y}$	$\bar{y}$
Total	Y	$\hat{Y}$
Razón	$R = Y/X = \bar{Y}/\bar{X}$	$\hat{R} = \bar{y}/\bar{x}$
Proporción	$P = A/N$	$p = a/n$

A número de elementos en la población que poseen la característica
a número de elementos en la muestra que poseen la característica

Ui unidad muestral i-ésima
X,Y,Z características o atributos que estamos interesados en observar en cada unidad Ui

Estimación de la media y el total poblacional

Propiedades de $\bar{y} = \frac{\sum y_i}{n}$ y $\hat{Y} = N\,\bar{y}$

1. Consistencia en el sentido de Cochran

si n=N $\rightarrow \bar{y}_n \equiv \bar{Y}$ y $\hat{Y} \equiv Y$

2. Estimadores insesgados $E\left(\bar{y}\right) = \bar{Y}$ y $E\left(\hat{Y}\right) = Y$ (probar)

3. Varianza de los estimadores

Notación (población):

$\sigma^2 = \sum^{N}\left(y_i - \bar{Y}\right)^2 \Big/ N$ varianza

$S^2 = \sum^{N}\left(y_i - \bar{Y}\right)^2 \Big/ (N-1)$ cuasivarianza

Estimación de la varianza de la media y el total poblacional

Pruebe que:

$$V(\bar{y}) = E\left(\bar{y} - \bar{Y}\right)^2 = \frac{S^2}{n}\frac{(N-n)}{N} = \frac{S^2}{n}(1-f) = \frac{\sigma^2}{n}\frac{(N-n)}{N-1}$$

ayuda

$$Cov(y_i, y_j) = -\sigma^2/(N-1) \qquad V\left(\bar{y}\right) = V\left(\frac{1}{n}\sum \bar{y}\right)$$

Error estándar de $\bar{y}$ es $\sigma_{\bar{y}} = \frac{S}{\sqrt{n}}\sqrt{1-f}$

Determine $V(\hat{Y})$ y el error estándar de $\hat{Y} = \sigma_{\hat{Y}}$

Corrección por población finita

En $V(\bar{y}) = \frac{\sigma^2}{n}\frac{(N-n)}{N}$; el término $\frac{(N-n)}{N}$ es el factor de corrección por poblaciones finitas

Si n/N<0.05 se puede ignorar f.c.p.f.

Pruebe que si X_i y Y_i son variables definidas en cada una de las unidades de la población

$$Cov\left(\bar{y}, \bar{x}\right) = \frac{N-n}{nN}\frac{1}{N-1}\sum\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right)$$

ayuda: crear variable auxiliar $u_i=y_i + x_i$ y determine $V(\bar{u})$

luego $V(\bar{u}) = E\left[\left(\bar{y} - \bar{Y}\right)\left(\bar{x} - \bar{X}\right)\right]^2$

desarrolle y elimine términos $V\left(\bar{y}\right)$ y $V\left(\bar{x}\right)$

Estimación del error estándar en la muestra

En el m.a.s. $E(s^2) = S^2$ donde $s^2 = \dfrac{\sum^{n}\left(y_i - \bar{y}\right)^2}{n-1}$

Ayuda: introduzca en s^2, $-\bar{Y}+\bar{Y}$ desarrolle y tome la esperanza

Las estimaciones insesgadas de las varianzas de $\bar{y}$ y de $\hat{Y}$ son:

$$V\left(\bar{y}\right) = \frac{s^2}{n}\left(1-f\right) \qquad V\left(\hat{Y}\right) = \frac{N^2 s^2}{n}\left(1-f\right)$$

y los errores estándar

$$s_{\bar{y}} = \frac{s}{\sqrt{n}}\sqrt{\left(1-f\right)} \qquad s_{\hat{Y}} = \frac{Ns}{\sqrt{n}}\sqrt{\left(1-f\right)}$$

El error estándar de la estimación sirve para

1. Comparar la presición obtenida en el m.a.s. con los de otros diseños muestrales
2. Estimar el tamaño de la muestra
3. Estimar la presición obtenida en la muestra seleccionada

Límites de confianza

Luego de obtenidas de la muestra las estimaciones de $\bar{y}$, $\hat{Y}$ y s se estiman intervalos de confianza.

¿Cómo se distribuyen $\bar{y}$ y $\hat{Y}$? Bajo condiciones muy generales podemos considerar que se distribuyen normal.

Luego los límites inferior L y superior U de los intervalos son:

$$\hat{Y}_L = N\bar{y} - t_\alpha \frac{Ns}{\sqrt{n}}\sqrt{1-f} \qquad \hat{Y}_U = N\bar{y} + t_\alpha \frac{Ns}{\sqrt{n}}\sqrt{1-f}$$

$$\hat{\bar{Y}}_L = \bar{y} - t_\alpha \frac{s}{\sqrt{n}}\sqrt{1-f} \qquad \hat{\bar{Y}}_U = \bar{y} + t_\alpha \frac{s}{\sqrt{n}}\sqrt{1-f}$$

Límites de confianza

t_α es el desvío de la distribución normal(0,1) correspondiente a la confianza asignada

$1-\alpha$	0.50	0.80	0.90	0.95	0.99
t_α	0.67	1.28	1.64	1.96	2.58

Si n<50 se puede usar la distribución t de student con n-1 g.l., pero el ajuste es bueno sii y_i~n y N$\rightarrow \infty$

Método alternativo de prueba:

Comprobación de los resultados del m.a.s. Cornfield (1944)

Creamos una v.a. dicotómica auxiliar a_i i=1,2,...,n

$a_i = 1$ si u_i esta en la muestra

$a_i = 0$ si u_i no esta en la muestra

así ahora $\bar{y} = \frac{1}{n}\sum^{N} a_i y_i$

de esta forma las a_i son v.a. y los y_i son números fijos. Determine ahora

$E\left(\bar{y}\right)$ y $V\left(\bar{y}\right)$

(Cochran pag. 54)

Muestreo aleatorio simple con restitución

Este diseño es igual que el anterior con la diferencia que una unidad cualquiera u_i puede aparecer en la muestra 1,2,...,n veces.

Determinación de la esperanza y varianza de

$$\bar{y} = \frac{1}{n}\sum^{n} y_i \qquad E\left(\bar{y}\right) = \bar{Y} \qquad V\left(\bar{y}\right) = \frac{\sigma^2}{n} = \frac{S^2}{n}\frac{\left(N-1\right)}{N}$$

Ayuda: Sea t_i una v.a. auxiliar que indica el número de veces que la i-esima unidad aparece en la muestra, t_i = 0,1,2,....,n i=1,2,...,N

Entonces las y_i son números fijos y asi:

$\bar{y} = \frac{1}{n}\sum^{N} t_i y_i$ t_i~b(n,p) p=1/N $E\left(t_i\right) = np = \frac{n}{N}$ $V\left(t_i\right) = npq = \frac{n}{N}\left(1-\frac{1}{N}\right)$ y

$Cov\left(t_i t_j\right) = -np_i p_j = -\frac{n}{N^2}$

Nota: Compare la $V\left(\bar{y}\right)$ en los m.a.s. con y sin reemplazo. Discútalo.

Muestreo para proporciones y porcentajes

Es muy frecuente que en estudios por muestreos deseemos estudiar el número total, proporción o el porcentaje de unidades que poseen alguna característica o atributo o caen en una clase predefinida.
Ejemplos:

La clasificación puede ser introducida en el cuestionario o en las tabulaciones.

Muestreo para proporciones y porcentajes

Notación:

Todas y cada una de las unidades caen en una de las posibles clases C y $\overline{C}$ (C=poseen la características y $\overline{C}$ =no poseen la característica).

	Población	Muestra
Nº de unidades en C	A	a
Proporción	$\frac{A}{N}=P$	$\frac{a}{n}=p$

Estimadores:

$$\hat{P}=p=\frac{a}{n} \qquad \hat{A}=Np=N\frac{a}{n}$$

¿Cómo se distribuye p en el m.a.s. (con y sin reemplazo) de manera exacta y de modo aproximado?
¿Binomial?, ¿Hipergeométrico?, ¿Normal?, ¿Otra?

Propiedades de los estimadores de P y de A:

Varianza de los estimadores

Demostrar:

(1) que p y Np son estimadores insesgados de P y de A respectivamente

(2) Que $S_p^2 = \frac{NPQ}{N-1}$ y $s_p^2 = \frac{npq}{n-1}$

(3) $V(P) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right)$ y $V\left(\hat{A}\right) = \frac{N^2 PQ}{n}\left(\frac{N-n}{N-1}\right)$

(4) Que $\hat{V}(p) = s_p^2 = \frac{N-n}{(n-1)N}pq$ es insesgada de V(P)

y $\hat{V}\left(\hat{A}\right) = \frac{N(N-n)}{(n-1)}pq$

(5) Qué sucede con los porcentajes?

Ayuda:

Crear una v.a. auxiliar y_i, $y_i=1$ si $u_i \in C$, $y_i=0$ si $u_i \notin C$ y aplicar lo obtenido en m.a.s. para variables continuas. Tener en cuenta que $\sum y_i^2 = \sum y_i$

Efecto de P en el error estándar:

Sabemos que $V(P) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right)$

si ignoramos el factor de corrección entonces

$V(P) = \frac{PQ}{n}$ si ignoramos n (es decir, n=1)

P	**0**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
PQ	0	900	1600	2100	2400	2500	2400	2100	1600	900	0
$\sqrt{PQ}$	0	30	40	46	49	50	49	46	40	30	0

PQ y $\sqrt{PQ}$ son simétricos y su máximo es en 50

Para $\sqrt{PQ} = 50$ se requiere n=100 para que $\sqrt{V(p)} = 5\%$

y n=2500 para que $\sqrt{V(p)} = 1\%$

Efecto de P en el error estándar (Continuación)

Si estamos interesados en estimar el número total de unidades A=NP, el anterior enfoque no es el adecuado (P = proporción)
¿Es posible que la estimación sea correcta módulo un error, digamos 5% del valor verdadero?
Es decir

$\frac{\sigma_{NP}}{NP}$; $\frac{\sigma_{NP}}{NP} = \frac{N}{\sqrt{n}} \frac{\sqrt{PQ}}{NP} \sqrt{\frac{N-1}{N-n}}$ es decir el coeficiente de variación de la estimación

si ignoramos el f.c.p.f. $CV = \sqrt{\frac{Q}{nP}}$ y n=1

P	0	0.1	0.5	1	5	10	20	30	40	50	60	70	80	90
$\sqrt{\frac{Q}{P}}$	∞	31.6	14.1	9.9	4.4	3.0	2.0	1.5	1.2	1.0	0.8	0.7	0.5	0.3

CV disminuye consistentemente al incremento de P, alto para CV<5%
Se requiere n grande si el atributo es raro.

HIPERGEOMÉTRICA:

En el m.a.s. sin reemplazo la distribución exacta de a (número de elementos de la muestra que poseen la característica) es una hipergeométrica.

$$\Pr\left(a, \overline{a} / A, \overline{A}\right) = \frac{\binom{A}{a}\binom{\overline{A}}{\overline{a}}}{\binom{N}{n}}$$

Nos da la Pr de que a=np, la distribución de p es tal que Pr(p)=Pr(a)

BINOMIAL:

Si A y $\overline{A} = N - A$ son suficientemente grande al tamaño de n, equivale a suponer que P es constante y en este

$$\Pr(a) = \binom{n}{a} P^a Q^{n-a}$$

NORMAL:

La distribución normal se puede utilizar como una aproximación a la distribución de p (convergencia) dependiendo fundamentalmente de la cantidad np. La siguiente tabla da los valores mínimos de np para hacer uso de la normal

P	Np (clase más pequeña)	n
0.5	15	30
0.4	20	50
0.3	24	80
0.2	40	200
0.1	60	600
0.05	70	1400
~ 0	80	∞

Límites de confianza:

Hipergeométrica

Queremos estimar por intervalos el valor de A, $\hat{A}_U$ el límite superior del intervalo es tal que la probabilidad de obtener **a** o menos individuos de C en la muestra es α_u $(\alpha_u = 0.025 \qquad 0.05)$

$$\alpha_U = \sum_{j=0}^{a} \Pr\left(j, n-j / \hat{A}_U, N - \hat{A}_U \right)$$

$\hat{A}_u$ se selecciona como el entero más pequeño que la satisface.

El límite inferior es de modo similar el entero más grande que la satisface:

$$\alpha_L = \sum_{j=0}^{a} \Pr\left(j, n-j / \hat{A}_L, N - \hat{A}_L \right)$$

así $\Pr\left(\hat{A}_L < A < \hat{A}_U \right) \leq 1 - (\alpha_U + \alpha_L)$ y para P $\hat{P}_U = \dfrac{\hat{U}_U}{N}$ y $\hat{P}_L = \dfrac{\hat{U}_L}{N}$

Límites de confianza: (continuación)

Normal

Los límites de confianza para P utilizando la aproximación a la normal

$$p \pm \left(t_\alpha \sqrt{1-f} \sqrt{\frac{pq}{n-1}} + \frac{1}{2n} \right)$$

donde $\hat{V}(p) = \frac{N-n}{N} \frac{pq}{n-1}$,

t_α es el desvío normal,

$f = \frac{n}{N}$

y $\frac{1}{2n}$ es un factor de corrección por continuidad.

Estimación del tamaño de la muestra:

¿Cómo determinar **n**?

1. Fijar el grado de precisión deseado en términos del error de estimación de acuerdo a los objetivos de la encuesta?
2. Encontrar una ecuación que relacione a **n** con la precisión deseada
3. Pre-estimar los valores de los parámetros de la ecuación que contiene a n.
4. Si se desean resultados válidos por dominio debe fijarse el error para cada uno y calcular el tamaño de muestra como suma de los tamaños de los dominios.
5. Cuando se miden varios atributos y se precisa el error para cada uno de ellos, esto origina varios tamaños de n, hay que reconciliarlos en uno sólo.
6. Buscar un compromiso entre la precisión (tamaño de n) y el costo.

Especificación de la precisión:

La precisión deseada en una encuesta se establece al definir la cantidad de error tolerable en las estimaciones obtenidas

$$\left(\left| \theta - \hat{\theta} \right| \leq e \right)$$

La precisión se debe fijar, en función de los objetivos y usos a los que se destinen los resultados de la investigación. En ocasiones le es difícil al encargado del estudio fijar la cantidad de error tolerable. El estadístico debe hacer un esfuerzo para explicar y dar elementos para ayudar a decidir el tamaño de la muestra.

Una tabla que muestre tamaño de error, muestra y costos alternos puede ser de mucha utilidad.

Tamaño de muestra para la proporción:

Sea **e** el margen máximo de error admisible en la estimación de P.

α el riesgo de que el error sea mayor a **e** ($1-\alpha$ el coeficiente de confianza)

N el tamaño de la población y σ_p la varianza de p (o su estimación)

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

conocemos que: $\sigma_p^2 = \frac{N-n}{N-1}\frac{PQ}{n}$

la formula que liga a **n** con la precisión deseada $e = t_{\alpha/2}\sqrt{\frac{N-n}{N-1}\frac{PQ}{n}}$

$t_{\alpha/2}$ absisa de la normal al resolver para **n**

$$n = \frac{Nt^2PQ}{e^2\left(N-1\right)+t^2PQ} \qquad \text{o equivalente} \qquad n = \frac{\frac{t^2PQ}{e^2}}{1+\frac{1}{N}\left(\frac{t^2PQ}{e^2}-1\right)} = \frac{n_o}{1+\frac{1}{N}\left(n_o-1\right)}$$

Como **P** es desconocida se sustituye por un estimación anticipada

Si **N** es grande $n_0 = \frac{t^2pq}{e^2} = \frac{pq}{V}$ donde $V = \frac{pq}{n}$ la varianza deseada

Tamaño de la muestra para el total A=NP con error relativo:

Al estimar el número total de unidades en la clase C podemos desear controlar el error relativo **r** en lugar del absoluto **e**. En este caso:

$$\Pr\left(\frac{|Np - NP|}{NP} \geq r\right) = \alpha \rightarrow \Pr\left(|p - P| \geq r\right) = \alpha$$

$$rP = t\sigma_p \qquad n = \frac{Nt^2Q}{r^2P(N-1)+t^2Q} \qquad n = \frac{\frac{t^2Q}{r^2P}}{1+\frac{1}{N}\left(\frac{t^2Q}{r^2P}-1\right)}$$

Cuando N es lo suficientemente grande

$$n_0 = \frac{t^2q}{r^2p} \qquad n = \frac{n_0}{1+(n_0-1)/N} \cong \frac{n_0}{1+(n_0/N)}$$

Tamaño de la muestra para la media y el total:

Al calcular el tamaño de la muestra para datos continuos se pueden presentar dos casos: (1) se desea controlar el error relativo **r** ó (2) el error es dado en términos absolutos de **e**

1. Se desea controlar el error relativo **r** , $0 < r < 1$

$$\Pr\left(\frac{\left|\bar{y}-\bar{Y}\right|}{\bar{Y}} \geq r\right) = \Pr\left(\frac{\left|N\,\bar{y}-N\,\bar{Y}\right|}{N\,\bar{Y}} \geq r\right) = \Pr\left(\left|\bar{y}-\bar{Y}\right| \geq r\,\bar{Y}\right) = \alpha \;;\quad 0 < \alpha < 1$$

como $\sigma_{\bar{Y}} = \sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}}$ $\qquad r\,\bar{Y} = t\sigma_{\bar{Y}} = t\sqrt{\frac{N-n}{N}}\frac{S}{\sqrt{n}}$

resolviendo para n

$$n = \frac{\left(\frac{tS}{r\,\bar{Y}}\right)^2}{1+\frac{1}{N}\left(\frac{tS}{r\,\bar{Y}}\right)^2} \quad \text{o} \quad n = \frac{n_0}{1+\frac{n_0}{N}} \quad \text{donde } n_0 = \frac{t^2S^2}{r^2\,\bar{Y}^2}$$

Tamaño de la muestra para la media y el total:

2. El error a controlar es **e**

$$\Pr\left(\left|\bar{y}-\bar{Y}\right| \geq e\right) = \alpha \qquad e = t\sigma_{\bar{Y}}$$

$$n = \frac{\left(\frac{tS}{e}\right)^2}{1+\frac{1}{N}\left(\frac{tS}{e}\right)^2} \qquad \text{o} \qquad n = \frac{n_0}{1+\frac{n_0}{N}}$$

donde $n_0 = \left(\frac{tS}{e}\right)^2 = \frac{S^2}{V}$,

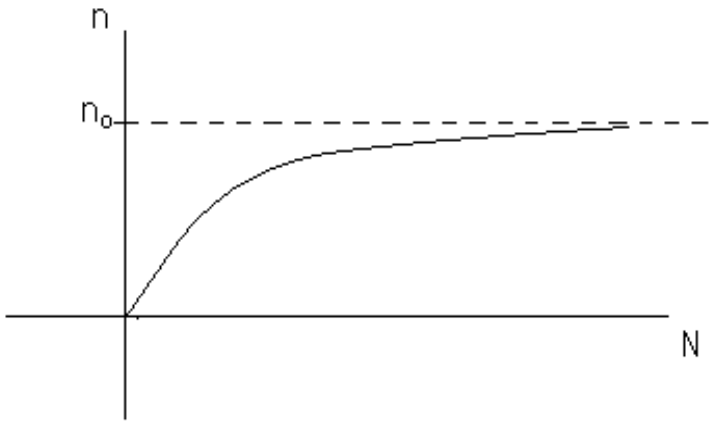
V es la varianza deseada de $\bar{y}$

Comportamiento de N:

Es de interés ver como es el comportamiento del tamaño de la muestra **n** en función de la población **N**. La función:

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

representa una hipérbole equilátera que pasa por el origen y tiene una asintota en n_o paralela al eje de las abscisas



Este resultado es importante pues nos dice que la misma precisión puede dar una muestra de tamaño **n** para una población de por ejemplo N_1=5000 elementos que para otro de N_2=100000. (siempre que verifique que $N_1 > n_o(n_o-1)$

Es de interés ver que n es inversamente proporcional al cuadrado del error absoluto y por lo tanto para, por ejemplo aumentar la precisión en el doble (error en la mitad) haría necesario aumentar la muestra en 4 veces.

Tamaño de la muestra para estudiar mas de una carcaterística:

En la mayoría de las encuestas se recoge información sobre más de una característica. En este caso el calculo de **n** es el siguiente:

1) Especificar los márgenes de error para las características mas importantes.
2) Estimar los tamaños de muestra para las características seleccionadas.
3) Tomar decisión sobre el valor de **n** de acuerdo a:
   a) Los **n** requeridos están suficientemente próximos.
   b) El **n** más grande está dentro de los límites del presupuesto
   c) Hay mucha variación entre los **n**, decidir por valores intermedios perdiendo precisión para algunas características.
4) Debe tomarse en cuenta que el tipo de diseño muestral puede mejorar la eficiencia o precisión de la muestra con **n** menores.

Estimación de la razón:

En algunas oportunidades es necesario estimar la razón entre dos variables apareadas en las unidades de la población donde se realizó el m.a.s.

La razón se define como: $R = \frac{\sum^{N} y_i}{\sum^{N} x_i}$ y su estimador: $\hat{R} = \frac{\sum^{n} y_i}{\sum^{n} x_i} = \frac{\hat{Y}}{\hat{X}} = \frac{\bar{y}}{\bar{x}}$

La distribución de $\hat{R}$ es complicada dada su estructura y la no independencia entre X e y.

En muestras pequeñas la distribución es asimetrica y $\hat{R}$ es por lo general ligeramente sesgada de R. En muestras grandes la distribución de $\hat{R}$ tiende a la normalidad y el sesgo es despreciable.

Pruebe que $ECM\left(\hat{R}\right) \doteq V\left(\hat{R}\right) \doteq \frac{1-f}{n\bar{X}} \frac{\sum \left(y_i - Rx_i\right)^2}{N-1}$

Ayuda: $\hat{R} - R \doteq \frac{y - R\bar{x}}{\bar{x}}$ es insesgado; en $E\left(y - R\bar{x}\right)^2$ haga $d_i = y_i - Rx_i$.

Defina $\bar{d}$ y $\bar{D} = 0$, $V(\bar{d})$ y $\frac{V\left(\bar{d}\right)}{n\bar{x}^2} = EMC\left(\hat{R}\right)$ $s\left(\hat{R}\right) = \frac{\sqrt{1-f}}{\sqrt{n}\,\bar{x}} \sqrt{\frac{\sum\left(y_i - \hat{R}\,x_i\right)}{n-1}}$

Estimación de la media en dominios de estudio

En muchos muestreos se requieren estimaciones en Dominios de Estudio
j-ésimo dominio; N_j unidades en la población; n_j unidades en la muestra
y_{jk} k-ésimo elemento del j-ésimo dominio,

$\sum_j N_j = N$ y $\sum_j n_j = n$; $\bar{y}_j = \frac{\sum_k y_{jk}}{n_j}$ n_j constantes de muestra en muestra

Si n_j y n son fijos: La probabilidad de sacar un conjunto específico de n_j unidades de las N_j del dominio j es $\frac{1}{\binom{N_j}{n_j}}$.

Luego si las n_j es constante de nuestra a muestra todo lo visto en el m.a.s. para $\bar{y}$ es aplicable a $\bar{y}_j$

$E(\bar{y}_j) = \bar{Y}_j$ es insesgado; $s_j^2 = \frac{\sum (y_{jk} - \bar{Y}_j)}{N_j 1}$

error estándar de $\bar{y}_j$ es $\frac{S_j}{n_j}\sqrt{1 - f_j}$ su estimación es $\frac{s_j}{n_j}\sqrt{1 - f_j}$

Estimación de totales en dominios de estudio:

Si estamos interesados en estimar totales de dominios de estudio $\bar{Y}$ se distinguen tres situaciones:

1) N_j conocido, entonces $\hat{Y} = N_j \bar{y}_j$ y su error estándar es N_j veces el de $\bar{y}_j$
2) Si conocemos el total Y podemos utilizar un estimador de razón

$\hat{Y}_j / \hat{Y}$ y multiplicarlo por Y.

3) Si no se conoce N_j ni Y se usa el estimador $\bar{Y}_j = \frac{N \sum y_{jk}}{n}$ el cual es insesgado

y su varianza es $\hat{\sigma}^2_{\hat{Y}_j} \frac{N^2 s'^2}{n}\left(1 - \frac{n}{N}\right)$ donde $s'^2 = \frac{1}{N-1}\left(\sum^{Nj} y_i - \frac{Y_j^2}{N}\right)$

Una estimación del error estádar de $\hat{Y}_j$ es: $s\left(Y_J\right) = \frac{Ns'}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$

Proporciones y totales sobre dominios de estudio:

Si queremos hacer estimaciones de proporciones o totales en dominios de una población observada por m.a.s. Los elementos de la muestra se clasifican así:

	Dominio 1	Dominio 2	.. .	Dominio i	.. .	Dominio k	Total
Clase	C C’	C C’		C C’		C C’	
Nº unidades	a_1 a'_1	a_2 a'_2		a_i a'_i		a_k a'_k	n
	n_1	n_2		n_i		n_k	n

$p_i=a_i/(a_i+a'_i)$

El intervalo de confianza (bajo normalidad) para p_i es

$$p_i \pm \left[t \sqrt{\left(1-\frac{n_i}{N_i}\right)\frac{p_i q_i}{(n_i-1)}+\frac{1}{n_i}} \right]$$

Proporciones y totales sobre dominios de estudio (continuación)

El total si N_i es el número de elementos en el dominio en el dominio i es: $\hat{A}_i = N_i p_i = \frac{N_i a_i}{a_i + a_i'}$

su error estándar $s(\hat{A}_i) = N\sqrt{1-(n_i/N_i)}\sqrt{p_i q_i/(n_i-1)}$

S N_i es conocido $\hat{A}_i^* = \frac{Na_i}{n}$ y $s(\hat{A}_i) = N\sqrt{1-n/N}\sqrt{pq/(n-1)}$

Como las proporciones en los dominios son estimados independientemente las comparaciones entre los dominios difrentes se puede hacer usando pruebas de contingencia.

Tamaño de la muestra para dominios de estudio

Para cada dominio de estudio tenemos que $n_i = s_i^2 / V$

y el tamaño de la muestra global será $n = \sum n_i$

Las s_i^2 individuales seran en promedio menores que s^2, pero a menudo solo un poco más, si hay k dominios $n \cong ks^2 / V$, en tanto para obtener estimaciones poblacionales $n = s^2 / V$

Ejemplo 1.

En el Centro Medico La Trinidad hay 484 pacientes que tienen deudas pendientes. De ellos se toma una m.a.s. de tamaño 9 y se obtienen los siguientes valores (expresados en miles de Bs)

i	1	2	3	4	5	6	7	8	9
y_i	33.50	32.00	52.00	43.00	40.00	41.00	45.00	42.50	39.00

a) Estimar $\overline{Y}$ y $\hat{Y}$, y determinar ambos intervalos de confianza para $\alpha = 0.05$ y $\alpha = 0.01$.

b) Calcular el tamaño de la muestra para $\overline{Y}$ suponiendo que $s^2 = 36$ con e=5 y e=10, y $\alpha = 0.05$ y $\alpha = 0.1$.

Y para $\hat{Y}$ con e=2000 y e=1500 y $\alpha = 0.05$ y $\alpha = 0.1$.

Respuesta:

$\sum y_i = 368$, $\sum y_i^2 = 15,332.5$, $\overline{y} = 368/9 = 40.89$,

$\hat{Y} = 484 * 40.89 = 19790.67$, $\hat{s}^2 = \frac{\sum (y_i - \overline{y})^2}{n-1} = \frac{\sum y_i^2 - (\sum y_i)^2 / n}{n-1} = 35.67$

$\hat{V}(\bar{y}) = \frac{(N-n)}{N}\frac{\hat{s}^2}{n} = \frac{(484-9)}{484}\frac{35.67}{9} = 3.89$, $\hat{s}_{\bar{y}} = 1.972$

$\hat{V}(\hat{Y}) = \hat{V}(N\bar{y}) = N^2\hat{V}(\bar{y}) = (484^2)3.89 = 911255.84$, $\hat{s}_{\hat{Y}} = 954.60$

Los intervalos de confianza son:

Para la media: $\bar{y} \mp t_\alpha \hat{s}_{\bar{y}}$

$\alpha = 0.05$ $t_\alpha = 1.96$ $40.89 \mp 1.96\,x1.972$ $(37.03;44.76)$

$\alpha = 0.01$ $t_\alpha = 2.58$ $40.89 \mp 2.58\,x1.972$ $(35.80;45.98)$

Para el total: $\hat{Y} \mp t_\alpha \hat{s}_{\hat{Y}}$

$\alpha = 0.05$ $t_\alpha = 1.96$ $19,790.76 \mp 1.96\,x954.60$ $(17,919.74;21,661.78)$

$\alpha = 0.01$ $t_\alpha = 2.58$ $19,790.76 \mp 2.58\,x954.60$ $(17,327.89;22,253.63)$

b) Tamaño de la muestra

$\hat{\bar{Y}}$; s^2=36 e=5 e=10 $\alpha = 0,05$ $t_\alpha = 1.96$ $\alpha = 0,01$ $t_\alpha = 2,58$

$n = \frac{n_0}{1+\frac{n_0}{N}}$ donde $n_0 = \frac{t^2 s^2}{e^2} = \frac{(1.96)^2(36)^2}{5^2} = 199$ $n_0 = \frac{(1.64)^2(36)^2}{5^2} = 140$

$$n_0 = \frac{(1.96)^2(36)^2}{10^2} = 50 \qquad n_0 = \frac{(1.64)^2(36)^2}{10^2} = 35$$

n_0	e=5	e=10
$t_\alpha = 1.64$	140	35
$t_\alpha = 1.96$	199	50

Ejemplo 2.

Un sondeo sobre una muestra de 2000 personas da como resultado de cara a las próximas elecciones:

No tienen intención de votar 800 personas

Si tienen intención de votar 1200 personas

Entre estos últimos la intención de voto es:

500 personas por A, 400 personas por B, 300 personas por C

a) Con una probabilidad del 95% estimar el porcentaje de abstención, de participación y los de cada una de las opciones A, B y C.

b) Suponiendo que la población total sea de 9,000,000 de votantes y utilizando la dispersión obtenida en la primera parte calcule el tamaño de muestra para estimar la abstención y el porcentaje de votantes que sufragarían por B y por A, tome e=5 y 10%, , $\alpha = 0.05$ y $\alpha = 0.01$. Calcule los costos de cada opción si cada encuesta es de Bs. 1,500 y los costos fijos son de 2,000,000 + 5% del costo de trabajo de campo.