Muestreo Estadístico

SEMANA-3

Raúl Alberto Pérez
raperez1@unal.edu.co
Profesor Asociado - Escuela de Estadística
Universidad Nacional de Colombia, Sede Medellín

Semestre 2021-I

Estimaciones en Subpoblaciones

En algunos casos es necesario realizar estimaciones sobre sub-grupos de la población.

La identificación de dichos sub-grupos, también llamados Dominios, se logra una vez la información de los elementos ha sido registrada.

En casos simples los dominios son excluyentes, es decir, un elementos sólo puede pertenecer a un dominio.

Estimación en Subpoblaciones o Dominios de Estudio

Caso de Estudio 2

Los siguientes datos corresponden a una M.A.S sin reemplazo de 40 familias de una comunidad de 4000 familias.
La variable de interés es el **ingreso mensual promedio** en salarios mínimos legales vigentes.
Además del ingreso, se registra el **sector** de la ciudad donde vive la familia.
Se desea estimar el **ingreso promedio** de las familias que viven en el **norte** (N) y de las que viven en el **sur** (S), así como el **ingreso total** en cada uno de esos sectores, con las estimaciones correspondientes de las varianzas de los estimadores utilizados.

Estimación en Subpoblaciones o Dominios de Estudio

Datos Caso de Estudio 2

Nro.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
V	N	S	S	N	S	N	N	S	N	N	N	N	S	S
I	2	6	3	2	5	7	9	8	7	4	3	4	10	5
Nro.	15	16	17	18	19	20	21	22	23	24	25	26	27	28
V	N	S	S	N	N	S	S	S	N	S	N	S	N	S
I	6	7	4	3	5	4	4	6	6	5	4	4	3	2
Nro.	29	30	31	32	33	34	35	36	37	38	39	40		
V	S	N	N	S	S	S	S	S	S	N	N	N		
I	2	4	5	5	12	11	4	7	8	4	5	3		

Denote por:

N_1- μ_1-τ_1- El **número total**, **ingreso promedio** e **ingreso total** de las familias que viven en el **norte**, respectivamente.

N_2- μ_2-τ_2- El **número total**, **ingreso promedio** e **ingreso total** de las familias que viven en el **sur**, respectivamente.

Se consideran dos escenarios:

1. **Se conoce el número total de familias que viven ya sea en el norte (N_1) o en el sur (N_2) de la ciudad.**
2. **Cuando NO se conocen.**

Estimación en Subpoblaciones o Dominios de Estudio

Caso de Estudio 3

Suponga que se desea realizar un estudio de muestreo en un Municipio A del departamento de Antioquia para estimar la **proporción de votantes** registrados con intención de voto por un candidato **X**. Para ello:

- Se dispone del **listado de todos los habitantes** mayores de edad que conforman el municipio, $N = 10000$ habitantes.
- Se opta por seleccionar una **muestra aleatoria simple** sin reemplazo de dicha población, $n = 500$.
- Se encuentra que en dicha muestra sólo **350 personas estaban inscritas para votar y de éstas 150 estaban a favor del candidato X.**

¿Cómo estimar la **proporción** de votantes registrados con intención de voto por el candidato **X?**

Estimación en Subpoblaciones o Dominios de Estudio

Caso de Estudio 4

Una multinacional desea abrir nuevos puestos de trabajo en un barrio de Medellín.

Para ello necesita estimar de las personas que **NO trabajan**, el **tiempo** (en meses) que el jefe de hogar ha completado sin trabajar.
La empresa cuenta con un listado de todos los hogares del barrio bajo estudio, conformado por 1000 hogares. Se decide:

- Seleccionar una **muestra piloto** de **10 hogares y se entrevista al jefe de hogar.**
- **Los datos que se obtuvieron fueron los siguientes:**

Hogar	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Trabaja (1:si, 0:no)	1	0	1	0	0	1	0	1	0	0
Tiempo en meses		3		12	5		7		8	2

¿Cómo estimar la **proporción** de hogares donde el jefe de hogar **NO** trabaja?

Anotaciones

1. Observar que en este caso, interesa hacer la estimación, no en la población completa definida mediante el marco de muestreo, sino en una Subpoblación (o dominio de estudio), para el tercer Caso de Estudio: Subpoblación de los votantes registrados, mientras que para el cuarto Caso de Estudio es: la subpoblación de los jefes de hogar que NO trabajan.
2. Para el Caso 3: $\hat{p}_1$: estimador de la **proporción de los votantes registrados** en el municipio A, que votarían por el candidato X=$\frac{a_1}{n_1}$.
   En esta situación n_1 es una variable aleatoria, y por tanto difiere del estimador visto inicialmente.
3. Para el Caso 4 $\hat{\mu}$: Estimador del **Tiempo promedio** en meses de los jefes de hogar sin trabajo= $\frac{\sum_i y_{ki}}{n_k}$, con y_{ki}: Tiempo en meses sin empleo del i-ésimo jefe de hogar, n_k: Número de jefes de hogar del total de los n-entrevistados que **NO** tenían empleo. En este caso n_k también es una **variable aleatoria**.

Estimación en Subpoblaciones o Dominios

Notaciones

- N: Número de unidades en la población
- N_k: Número de unidades que pertenecen a la subpoblación k.
- y_{ki}: Valor de la variable de interés para la i-ésima unidad en la k-ésima subpoblación.
- τ_k: Total de la subpoblación: $\sum_{i=1}^{N_k} y_{ki}$.
- $\mu_k = \frac{\tau_k}{N_k}$: Media de la subpoblación.

Se selecciona una M.A.S. de n-unidades de las N-unidades de la población. Adicionalmente:
n_k : número de unidades de la muestra que hacen parte de la subpoblación k.

Estimación de la Media en una Subpoblación

La media muestral de las n_k-unidades:

$$\overline{y}_k = \frac{\sum_{i=1}^{n_k} y_{ki}}{n_k},$$

es un estimador insesgado de la media poblacional μ_k de la variable de interés en la k-ésima subpoblación.

En efecto:

Dado n_k, cualquier combinación posible de n_k unidades de las N_k unidades de la subpoblación tiene igual probabilidad de estar incluída en la muestra.

Luego:

$$E\left[\bar{y}_k | n_k\right] = \mu_k,$$

Aplicando la propiedad **iterativa de la esperanza condicional** se tiene:

$$E\left[\bar{y}_k\right] = E\left[E\left(\bar{y}_k | n_k\right)\right] = E[\mu_k] = \mu_k.$$

Estimación de la Media en una Subpoblación

Adicionalmente, la **Varianza:**

$$V(\bar{y}_k) = Var\left(E\left[\bar{y}_k|n_k\right]\right) + E\left[Var\left(\bar{y}_k|n_k\right)\right] = Var(\mu_k) + E\left[\frac{N_k - n_k}{N_k}\frac{\sigma_k^2}{n_k}\right] = E\left[\left(\frac{1}{n_k} - \frac{1}{N_k}\right)\sigma_k^2\right]$$

$$= \sigma_k^2\left\{E\left[\frac{1}{n_k}\right] - \frac{1}{N_k}\right\},$$

donde $\sigma_k^2 := \frac{1}{N_k - 1}\sum_{i=1}^{N_k}(y_{ki} - \mu_k)^2$.

NOTA:

En este caso n_k es una v.a Hipergeométrica (N, N_k, n, n_k), donde su media es: $E[n_k] = n\frac{N_k}{N}$, y al aplicar la desigualdad de Jensen:

$$E\left[\frac{1}{n_k}\right] > \frac{1}{E[n_k]} = \frac{N}{nN_k}.$$

De la expresión anterior, se tiene que la varianza del estimador de la media es mayor que la varianza del estimador de una M.A.S de tamaño fijo $n\frac{N_k}{N}$.

Estimación de la Media en una Subpoblación

La varianza estimada de dicho estimador está dada por:

$$\widehat{var}[\overline{y}_k] = \left(\frac{N_k - n_k}{N_k}\right)\left(\frac{S_k^2}{n_k}\right) \quad , \quad \text{si} \quad N_k - \text{es conocida} \quad , \quad \text{y}$$

$$\widehat{var}[\overline{y}_k] = \left(\frac{N - n}{N}\right)\left(\frac{S_k^2}{n_k}\right) \quad , \quad N_k - \text{es desconocida}, \quad \hat{N}_k = N(n_k/n)$$

con S_k^2-la varianza de los elementos de la muestra pertenecientes a la k-ésima subpoblación, ie.

$$S_k^2 = \frac{\sum_{i=1}^{n_k}(y_{ki} - \overline{y}_k)^2}{n_k - 1}$$

Estimación de Totales Poblacionales en Subpoblaciones

Si N_k-se conoce, entonces un estimador insesgado de τ_k es:

$$\hat{\tau}_k = N_k \overline{y}_k$$

con varianza estimada dada por:

$$\widehat{var}[\hat{\tau}_k] = N_k^2 \widehat{var}[\overline{y}_k]$$

Si N_k-es desconocido, entonces un estimador insesgado de τ_k es:

$$\hat{\tau}_k^\star = \left(\frac{N}{n}\right)\sum_{i=1}^{n_k} y_{ki} = \left(\frac{N}{n}\right)\sum_{i=1}^{n} y_{i^\star} = N\overline{y}^\star,$$

con varianza estimada dada por:

$$\widehat{var}[\hat{\tau}_k^\star] = N^2\widehat{var}[\overline{y}^\star] = N^2\left(\frac{N-n}{N}\right)\left(\frac{S^{\star 2}}{n}\right),$$

donde,

$$S^{\star 2} = \frac{\sum_{i=1}^{n}(y_i^\star - \overline{y}^\star)^2}{n-1} \quad \text{y} \quad \overline{y}^\star = \frac{\sum_{i=1}^{n} y_i^\star}{n},$$

con $Y^\star$-una nueva variable de interés que es idéntica a Y, para todo elemento que pertenezca a la subpoblación de interés y es cero-en otro caso.

Intervalos de Confianza Para μ_k

Un intervalo de confianza del $(1-\alpha)\times 100\%$ para μ_k está dado por:

$$\bar{y}_k \pm t_{1-\alpha/2,n_k-1}\sqrt{\hat{V}(\bar{y}_k)},$$

Estiamción de I.C para p_k

En forma análoga se hace para la estimación de la **proporción** en la subpoblación k: $\hat{p}_k = \frac{a_k}{n_k}$, con a_k: Número de unidades en la muestra que **cumplen el atributo** en la muestra y que pertenecen a la subpoblación k.

$$\hat{p}_k \pm t_{1-\alpha/2,n_k-1}\sqrt{\hat{V}(\bar{p}_k)},$$

donde: $\hat{V}(\hat{p}_k) = \left(\frac{N_k - n_k}{N_k}\right)\left(\frac{\hat{p}_k(1-\hat{p}_k)}{n_k - 1}\right).$

EJEMPLO: Considere la siguiente información que corresponde a una M.A.S de 40-familias de una comunidad de 4000 familias, donde la variable de interés es el ingreso mensual medio en número de salarios mínimos. Además del ingreso, se registra el sector de la ciudad donde vive la familia. Se desea estimar el ingreso promedio de las familias que viven en el norte (N) y de las que viven en el sur (S), así como el ingreso total en cada uno de esos sectores, con las estimaciones correspondientes de las varianzas de los estimadores utilizados.

Nro.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
V	N	S	S	N	S	N	N	S	N	N	N	N	S	S
I	2	6	3	2	5	7	9	8	7	4	3	4	10	5
Nro.	15	16	17	18	19	20	21	22	23	24	25	26	27	28
V	N	S	S	N	N	S	S	S	N	S	N	S	N	S
I	6	7	4	3	5	4	4	6	6	5	4	4	3	2
Nro.	29	30	31	32	33	34	35	36	37	38	39	40		
V	S	N	N	S	S	S	S	S	S	N	N	N		
I	2	4	5	5	12	11	4	7	8	4	5	3		

Sean las siguientes cantidades:

N_1-El número total de familias que viven en el norte.

N_2-El número total de familias que viven en el sur.

μ_1-El ingreso promedio de familias que viven en el norte.

μ_2-El ingreso promedio de familias que viven en el sur.

τ_1-El ingreso total de familias que viven en el norte.

τ_2-El ingreso total de familias que viven en el sur.

Se tomarán en cuenta las dos situaciones diferentes: Cuando se conoce el número total de familias que viven ya sea en el norte (N_1)o en el sur (N_2) de la ciudad y cuando no se conocen.

Caso-I: Cuando se conocen los tamaños de las subpoblaciones.

$$\text{Supóngase que:} \quad N_1 = 1600 \quad \text{y} \quad N_2 = 2400.$$

De la tabla de datos se observa que:

$$n_1 = 19 \quad \text{y} \quad n_2 = 21.$$

Las estimaciones para los promedios con sus respectivas varianzas son:

Para la primera subpoblación son:

$$\hat{\mu}_1 = \overline{y}_1 = \frac{(2+2+7+\cdots+4+5+3)}{19} = \frac{86}{19} = 4,53\,, \text{ sal mín por fam}$$

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \overline{y}_1)^2}{n_1 - 1} = \frac{\sum_{i=1}^{19}(y_{1i} - 4.53)^2}{18} = 3.3743,$$

su varianza estimada es:

$$var[\overline{y}_1] = \frac{N_1 - n_1}{N_1}\left(\frac{S_1^2}{n_1}\right) = \frac{1600 - 19}{1600}\left(\frac{3.3743}{19}\right) = 0.1755.$$

Para la segunda subpoblación son:

$$\hat{\mu}_2 = \overline{y}_2 = \frac{(6 + 3 + 5 + \cdots + 4 + 7 + 8)}{21} = 5.81\,, \text{ sal mín por fam}$$

$$S_2^2 = \frac{\sum_{i=1}^{n_2}(y_{2i} - \overline{y}_2)^2}{n_2 - 1} = \frac{\sum_{i=1}^{21}(y_{2i} - 5.81)^2}{20} = 7.5619,$$

su varianza estimada es:

$$var[\overline{y}_2] = \frac{N_2 - n_2}{N_2}\left(\frac{S_2^2}{n_2}\right) = \frac{1600 - 21}{2400}\left(\frac{7.5619}{21}\right) = 0.3569.$$

Las estimaciones para los totales con sus respectivas varianzas son:

Para la primera subpoblación son:

$$\hat{\tau}_1 = N_1 \overline{y}_1 = 1600(4,53) = 7248 \ , \quad \text{salarios mínimos.}$$

su varianza estimada es:

$$var[\hat{\tau}_1] = N_1^2 var[\overline{y}_1] = (1600)^2(0.1755) = \boxed{449280}$$

Para la segunda subpoblación son:

$$\hat{\tau}_2 = N_2 \overline{y}_2 = 2400(5.81) = 13944 \ , \quad \text{salarios mínimos.}$$

su varianza estimada es:

$$var[\hat{\tau}_2] = N_2^2 var[\overline{y}_2] = (2400)^2(0.3569) = \boxed{2055744}$$

Caso-II: Cuando los tamaños de las subpoblaciones son desconocidos.

Las estimaciones para los promedios con sus respectivas varianzas son:

Para la primera subpoblación son:

$$\hat{\mu}_1 = \overline{y}_1 = \frac{(2+2+7+\cdots+4+5+3)}{19} = 4,53 \text{ , sal mín por fam}$$

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(y_{1i}-\overline{y}_1)^2}{n_1-1} = \frac{\sum_{i=1}^{19}(y_{1i}-4.53)^2}{18} = 3.3743$$

su varianza estimada es:

$$var[\overline{y}_1] = \frac{N-n}{N}\left(\frac{S_1^2}{n_1}\right) = \frac{4000-40}{4000}\left(\frac{3.3743}{19}\right) = \boxed{0.1758}$$

Para la segunda subpoblación son:

$$\hat{\mu}_2 = \overline{y}_2 = \frac{(6+3+5+\cdots+4+7+8)}{21} = 5.81\ , \text{ sal mín por fam}$$

$$S_2^2 = \frac{\sum_{i=1}^{n_2}(y_{2i}-\overline{y}_2)^2}{n_2-1} = \frac{\sum_{i=1}^{21}(y_{2i}-5.81)^2}{20} = 7.5619$$

su varianza estimada es:

$$var[\overline{y}_2] = \frac{N-n}{Nn_2}s_2^2 = \frac{(4000-40)}{4000(21)}(7.5619) = \boxed{0.3565}$$

Para las estimaciones de los totales en este caso se procede como sigue:

Para la primera subpoblación son:

$$\hat{\tau}_1^\star = \frac{N}{n}\sum_{i=1}^{n_k} y_{ki} = N\bar{y}^\star = \frac{4000}{40}(2+2+7+\cdots+4+5+3) = 8600$$

con varianza estimada dada por:

$$var[\hat{\tau}_1^\star] = N^2\left(\frac{N-n}{N}\right)\left(\frac{S^{\star 2}}{n}\right) = (4000)^2\left(\frac{4000-40}{4000}\right)\left(\frac{6.7974}{40}\right) = 2691770$$

donde,

$$S^{\star 2} = \frac{\sum_{i=1}^{n}(y_i^\star - \overline{y}^\star)^2}{n-1} = 6.7974 \quad , \quad \overline{y}^\star = \frac{\sum_{i=1}^{n} y_i^\star}{n} = 2.15$$

Para estimar $S^{\star 2}$ y a $\overline{y}^\star$, se define la nueva variable de interés $Y^\star$ de la siguiente forma:

$Y^{\star}$-es idéntica a Y, para todo elemento que pertenezca a la subpoblación número uno y es cero-en otro caso, es decir:

$$y_1^{'} = 2 \quad y_2^{'} = 0 \quad y_3^{'} = 0 \quad y_4^{'} = 2 \quad y_5^{'} = 0 \quad y_6^{'} = 7 \quad y_7^{'} = 9 \quad y_8^{'} = 0$$

$$y_9^{'} = 7 \quad y_{10}^{'} = 4 \quad y_{11}^{'} = 3 \quad y_{12}^{'} = 4 \quad y_{13}^{'} = 0 \quad y_{14}^{'} = 0 \quad y_{15}^{'} = 6 \quad y_{16}^{'} = 0$$

$$y_{17}^{'} = 0 \quad y_{18}^{'} = 3 \quad y_{19}^{'} = 5 \quad y_{20}^{'} = 0 \quad y_{21}^{'} = 0 \quad y_{22}^{'} = 0 \quad y_{23}^{'} = 6 \quad y_{24}^{'} = 0$$

$$y_{25}^{'} = 4 \quad y_{26}^{'} = 0 \quad y_{27}^{'} = 3 \quad y_{28}^{'} = 0 \quad y_{29}^{'} = 0 \quad y_{30}^{'} = 4 \quad y_{31}^{'} = 5 \quad y_{32}^{'} = 0$$

$$y_{33}^{'} = 0 \quad y_{34}^{'} = 0 \quad y_{35}^{'} = 0 \quad y_{36}^{'} = 0 \quad y_{37}^{'} = 0 \quad y_{38}^{'} = 4 \quad y_{39}^{'} = 5 \quad y_{40}^{'} = 3$$

Para la segunda subpoblación son:

$$\hat{\tau}_2^{\star} = \frac{N}{n}\sum_{i=1}^{n_k} y_{ki} = \frac{4000}{40}(6+3+5+\cdots+4+7+8) = 1200 \text{ , salarios mínimos}$$

con varianza estimada dada por:

$$var[\hat{\tau}_2^{\star}] = N^2\left(\frac{N-n}{N}\right)\left(\frac{S^{\star 2}}{n}\right) = (4000)^2\left(\frac{4000-40}{4000}\right)\left(\frac{12.5103}{40}\right) = \boxed{4954078}$$

donde,

$$S^{\star 2} = \frac{\sum_{i=1}^{n}(y_i^{\star}-\overline{y}^{\star})^2}{n-1} = 12.5103 \quad , \quad \overline{y}^{\star} = \frac{\sum_{i=1}^{n} y_i^{\star}}{n} = 3.05$$

donde nuevamente para estimar a $S^{\star 2}$ y a $\overline{y}^{\star}$, se define la nueva variable de interés $Y^{\star}$ de la siguiente forma:

$Y^{\star}$-es idéntica a Y, para todo elemento que pertenezca a la subpoblación número dos y es cero-en otro caso, es decir:

$$y_1^{'} = 0 \quad y_2^{'} = 6 \quad y_3^{'} = 3 \quad y_4^{'} = 0 \quad y_5^{'} = 5 \quad y_6^{'} = 0 \quad y_7^{'} = 0 \quad y_8^{'} = 8$$

$$y_9^{'} = 0 \quad y_{10}^{'} = 0 \quad y_{11}^{'} = 0 \quad y_{12}^{'} = 0 \quad y_{13}^{'} = 10 \quad y_{14}^{'} = 5 \quad y_{15}^{'} = 0 \quad y_{16}^{'} = 7$$

$$y_{17}^{'} = 4 \quad y_{18}^{'} = 0 \quad y_{19}^{'} = 0 \quad y_{20}^{'} = 4 \quad y_{21}^{'} = 4 \quad y_{22}^{'} = 6 \quad y_{23}^{'} = 0 \quad y_{24}^{'} = 5$$

$$y_{25}^{'} = 0 \quad y_{26}^{'} = 4 \quad y_{27}^{'} = 0 \quad y_{28}^{'} = 2 \quad y_{29}^{'} = 2 \quad y_{30}^{'} = 0 \quad y_{31}^{'} = 0 \quad y_{32}^{'} = 5$$

$$y_{33}^{'} = 12 \quad y_{34}^{'} = 11 \quad y_{35}^{'} = 4 \quad y_{36}^{'} = 7 \quad y_{37}^{'} = 8 \quad y_{38}^{'} = 0 \quad y_{39}^{'} = 0 \quad y_{40}^{'} = 0$$

En los resultados anteriores se puede observar la gran contribución que tiene el conocimiento previo de los tamaños de las subpoblaciones sobre la disminución de las varianzas de las estimaciones de los totales.

En conclusión, siempre que los tamaños de las subpoblaciones se conozcan, estos deben usarse en el proceso de estimación.

Estimación de Proporciones y Totales de elementos con una cierta característica de interés en Subpoblaciones

Para este caso se define la siguiente variable indicadora:

$$y_{ki} = \begin{cases} 1 & - \quad \text{Si el i-ésimo elemento de la } k\text{-ésima subpoblación posee el atributo de interés} \\ 0 & - \quad e.o.c \end{cases}$$

Sean lo siguientes elementos: A_k-el número de elementos de la k-ésima subpoblación con el atributo de interés.

a_k-el número de elementos en la muestra correspondientes a la k-ésima subpoblación con el atributo de interés.

$P_k = \frac{A_k}{N_k}$-proporción de elementos de la k-ésima subpoblación con el atributo de interés.

$p_k = \frac{a_k}{n_k}$-proporción muestral de elementos de la k-ésima subpoblación con el atributo de interés.

Teorema: $p_k = \frac{a_k}{n_k}$-es un estimador insesgado de P_k, con varianza estimada dada por:

$$var[p_k] = \left(\frac{N_k - n_k}{N_k}\right)\left(\frac{p_k q_k}{n_k - 1}\right)$$

si N_k-conocido y $q_k = 1 - p_k$.

Si N_k-es desconocido, entonces se tiene que:

$$var[p_k] = \left(\frac{N - n}{N}\right)\left(\frac{p_k q_k}{n_k - 1}\right), \qquad \hat{N}_k = N(n_k/n)$$

Similarmente, para el total de elementos en la subpoblación con cierta característica de interés, se procede como sigue:

Si N_k-es conocido ,

$$\hat{A}_k = N_k p_k \ , \quad \text{y} \qquad var[\hat{A}_k] = N_k^2 var[p_k]$$

y si N_k-es desconocido se tiene lo siguiente:

$$\hat{A}_k^\star = \frac{N}{n} a_k = N\left(\frac{a_k}{n}\right) = Np^\star \ , \quad \text{y}$$

$$var[\hat{A}_k^\star] = N^2 var[p^\star] = N^2\left(\frac{N-n}{n}\right)\frac{p^\star q^\star}{n-1}$$

con,

$$p^\star = \frac{a_k}{n} \quad ; \quad q^\star = 1 - p^\star$$

EJEMPLO: En una m.a.s sin reemplazo de 400 personas de una población de 20000, además de registrar el nivel educativo (estudios universitarios $k = 1$, estudios secundarios $k = 2$, estudios primarios o ninguno $k = 3$), se preguntó a las personas si tenían intenciones de votar en las próximas elecciones ($y_{ki} = 1$, si la i-ésima persona dijo que sí, $y_{ki} = 0$, si la i-ésima persona dijo que no). Los resultados muestrales fueron los siguientes:

$$n_1 = 100 \ , \quad a_1 = 80$$

$$n_2 = 180 \ , \quad a_2 = 108$$

$$n_3 = 120 \ , \quad a_3 = 48$$

Suponga que inicialmente se conocen los tamaños de las subpoblaciones y que son:

$$N_1 = 4000\ , \quad N_2 = 7000 \quad \text{y} \quad N_3 = 9000$$

Con base en esta información suministrada, las estimaciones para las proporciones (o porcentajes) con sus respectivas varianzas son:

$$p_1 = \frac{a_1}{n_1} = \frac{80}{100} = 0.8$$

$$var[p_1] = \frac{N_1 - n_1}{N_1}\frac{p_1 q_1}{n_1 - 1} = \frac{(4000 - 100)}{4000}\frac{(0.8)(0.2)}{100 - 1} = \boxed{0.001576}$$

$$p_2 = \frac{a_2}{n_2} = \frac{108}{180} = 0.6$$

$$var[p_2] = \frac{N_2 - n_2}{N_2} \frac{p_2 q_2}{n_2 - 1} = \frac{(7000 - 180)}{7000} \frac{(0.6)(0.4)}{180 - 1} = \boxed{0.001306}$$

$$p_3 = \frac{a_3}{n_3} = \frac{48}{120} = 0.4$$

$$var[p_3] = \frac{N_3 - n_3}{N_3} \frac{p_3 q_3}{n_3 - 1} = \frac{(9000 - 120)}{9000} \frac{(0.4)(0.6)}{120 - 1} = \boxed{0.001990}$$

Ahora, las estimaciones del total de personas con intensiones de votar en cada una de las tres subpoblaciones, con sus respectivas varianzas estimadas, son:

$$\hat{A}_1 = N_1 p_1 = 4000(0.8) = 3200$$

$$var[\hat{A}_1] = N_1^2 \left(\frac{N_1 - n_1}{N_1}\right) \frac{p_1 q_1}{n_1 - 1} = 4000(4000 - 100)\frac{(0.8)(0.2)}{100 - 1} = 25216$$

$$\hat{A}_2 = N_2 p_2 = 7000(0.6) = 4200$$

$$var[\hat{A}_2] = N_2^2 \left(\frac{N_2 - n_2}{N_2}\right) \frac{p_2 q_2}{n_2 - 1} = 7000(7000 - 180)\frac{(0.6)(0.4)}{180 - 1} = 64008$$

$$\hat{A}_3 = N_3 p_3 = 9000(0.4) = 3600$$

$$var[\hat{A}_3] = N_3^2 \left(\frac{N_3 - n_3}{N_3}\right) \frac{p_3 q_3}{n_3 - 1} = 9000(9000 - 120)\frac{(0.4)(0.6)}{120 - 1} = 161183$$

Ahora suponga que los tamaños de las subpoblaciones son desconocidos:

$$p_1 = \frac{a_1}{n_1} = \frac{80}{100} = 0.8$$

$$var[p_1] = \frac{N-n}{N}\left(\frac{p_1 q_1}{n_1 - 1}\right) = \frac{20000-400}{20000}\frac{(0.8)(0.2)}{100-1} = 0.001584$$

$$p_2 = \frac{a_2}{n_2} = \frac{108}{180} = 0.6$$

$$var[p_2] = \frac{N-n}{N}\left(\frac{p_2 q_2}{n_2 - 1}\right) = \frac{20000-400}{20000}\frac{(0.6)(0.4)}{180-1} = 0.001314$$

$$p_3 = \frac{a_3}{n_3} = \frac{48}{120} = 0.4$$

$$var[p_3] = \frac{N-n}{N}\left(\frac{p_3 q_3}{n_3 - 1}\right) = \frac{20000-400}{20000}\frac{(0.4)(0.6)}{120-1} = 0.001976$$

Ahora, las estimaciones del total de personas con intensiones de votar en cada una de las tres subpoblaciones, con sus respectivas varianzas estimadas, son:

$$\hat{A}_1^\star = \frac{N}{n} a_1 = \frac{20000}{400} 80 = 4000 = Np^\star = 20000(0.2)$$

$$var[\hat{A}_1^\star] = N^2 \left(\frac{N-n}{N}\right)\left(\frac{p_1^\star q_1^\star}{n-1}\right) = 20000(20000-400)\frac{(0.2)(0.8)}{400-1} = \boxed{157193}$$

$p^\star = \frac{a_1}{n} = \frac{80}{400} = 0.2$

$$\hat{A}_2^\star = \frac{N}{n} a_2 = \frac{20000}{400} 108 = 5400 = 4000 = Np^\star = 20000(0.27)$$

$$var[\hat{A}_2^\star] = N^2 \left(\frac{N-n}{N}\right)\left(\frac{p_2^\star q_2^\star}{n-1}\right) = 20000(20000-400)\frac{(0.27)(0.73)}{400-1} = \boxed{193642}$$

$p^\star = \frac{a_2}{n} = \frac{108}{400} = 0.27$

$$\hat{A}_3^{\star} = \frac{N}{n} a_3 = \frac{20000}{400} 48 = 24004000 = Np^{\star} = 20000(0.12)$$

$$var[\hat{A}_3^{\star}] = N^2 \left(\frac{N-n}{N}\right)\left(\frac{p_3^{\star} q_3^{\star}}{n-1}\right) = 20000(20000-400)\frac{(0.12)(0.88)}{400-1} = 103747$$

$p^{\star} = \frac{a_3}{n} = \frac{48}{400} = 0.12$

Nuevamente se observan cambios en las varianzas de los totales de elementos con la característica de interés, cuando no se conocen los tamaños de las subpoblaciones.

NOTA: Todas las varianzas estimadas anteriormente son útiles para establecer I.C basados en la distribución normal, cuando los tamaños de muestras son suficientemente grandes, de lo contrario hay que usar la aproximación binomial.

Tarea 1 fecha Entrega: HASTA 19 de Marzo al inicio de Clase

1. Considere el Caso de Estudio 3, diapositiva 5. Responda la pregunta de interés, reporte e interprete el respectivo intervalo de confianza del 95%.
2. Considere el Caso de Estudio 4, diapositiva 6. Responda las dos preguntas de interés, reporte e interprete los respectivos intervalos de confianza del 95%.
3. Del Taller 1 de MAS, enviado, resuelva el problema asignado. Para ello, en la siguiente tabla aparece, según el número asignado en la lista de estudiantes adjunta (**NLista) el problema a resolver asignado (Pr. As.)**

   ALEATORIAMENTE.

NLista	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Pr. As.	12	18	15	8	4	6	5	11	20	2	13	19	17	1	1

NLista	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Pr. As.	12	18	15	8	4	6	16	3	5	11	20	2	13	19	14

NLista	31	32	33	34	35	36	37	38	39	40
Pr. As.	9	10	12	3	10	15	20	7	19	6

	APELLIDOS Y NOMBRE		APELLIDOS Y NOMBRE
1	Agudelo Carmona, Valentina	21	Murillo Anzola, Christian Camilo
2	Alvarez Morales, Guillermo	22	Naranjo Garcia, Yised Katerine
3	Aristizabal Echeverri, Genaro Alfonso	23	Nieto Morales, Alexis Andrés
4	Arzuaga Gonzalez, Maria Isabel	24	Pabón Palacio, Antonio
5	Bula Charry, Valentina	25	Palacios Duque, Sara
6	Bula Isaza, Juan Daniel	26	Pico Arredondo, Daniela
7	Carvajal Torres, Santiago	27	Rios Castro, Kaline Andrea
8	Duque Calle, David	28	Rios Garcia, Jhon Alexander
9	Franco Valencia, Santiago	29	Rojas Bolaños, Carlos Arturo
10	Galeano Arenas, Juan José	30	Salazar Mejía, Alejandro
11	Galeano Muñoz, Simón Pedro	31	Serrano Santos, Andrea
12	Garcia Muñoz, Jhonatan Smith	32	Suarez Ledesma, Jose Luis
13	Gaviria Sanchez, Sebastian	33	Tous Diaz, Cleidy Jimena
14	Gómez Valencia, Beatriz Valentina	34	Vanegas Castaño, Valentina
15	Granada Alvarez, Santiago	35	Vasco Ruiz, Daniela
16	Henao Vargas, Luisa Fernanda	36	Vásquez Gómez, Kleider Stiven
17	Hoyos Arias, John Daniel	37	Velez Rivera, Vanessa
18	Hoyos Peña, Laura Daniela	38	Yepes Pareja, Maria Fernanda
19	Jaramillo Calle, David	39	Zabaleta Cardeño, Carmen Daniela
20	Martínez Echavarría, Juan Pablo	40	Zuluaga Ayala, Santiago