

# Estimación en dos ocasiones usando la técnica de conteo de ítem de Hussain para poblaciones finitas

Andrés Cruz Pérez, M.Sc  
Luz Mery Gonzalez, Ph.D.

Universidad Santo Tomas  
Universidad Nacional de Colombia

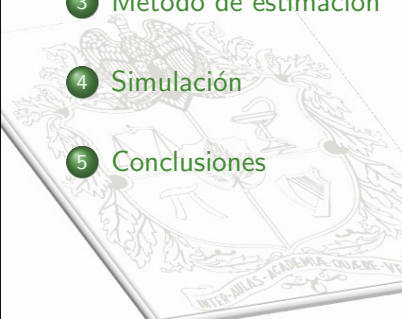
15 de julio de 2019



UNIVERSIDAD SANTO TOMÁS  
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

# Contenido

- 1 Introducción
- 2 Marco Teórico
- 3 Método de estimación
- 4 Simulación
- 5 Conclusiones



UNIVERSIDAD SANTO TOMÁS  
PRIMER CUAESTRO UNIVERSITARIO DE COLOMBIA

## 1 Introducción



UNIVERSIDAD SANTO TOMAS  
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

# Motivación

¿Cómo podemos obtener respuestas veraces a preguntas sensibles y medir su evolución a lo largo del tiempo?

Ejemplos:

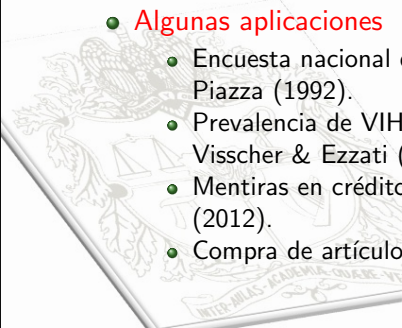
- Determinar el aumento o disminución en la evasión de impuestos.
- Estudios mensuales en los que se mide el uso de sustancias alucinógenas en una población de adolescentes.
- Estudios de opinión política conducidos en intervalos regulares para medir la prevalencia del racismo.
- Prevalencia del fraude académico en exámenes de la universidad.



UNIVERSIDAD SANTO TOMÁS  
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

# Contextualización

- **Desafío de medir actitudes y comportamientos sensibles**
  - Sesgo por deseabilidad social.
  - Sesgo por no respuesta.
- **Proteger la privacidad a través de preguntas indirectas**
  - Warner (1965) Método de respuesta aleatorizada.
  - Miller (1984) Técnica de conteo de ítem.
- **Algunas aplicaciones**
  - Encuesta nacional de raza y política. Sniderman, Tetlock & Piazza (1992).
  - Prevalencia de VIH. Droitcour, Caspar, Hubbard, Parsley, Visscher & Ezzati (1991).
  - Mentiras en créditos de libre inversión. Karlan & Zinman (2012).
  - Compra de artículos robados. (Kuha & Jackson 2014).



# Contenido

- 1 Introducción
- 2 Marco Teórico
- 3 Método de estimación
- 4 Supuestos
- 5 Conclusiones



UNIVERSIDAD SANTO TOMÁS  
PRIMER CUAESTRO UNIVERSITARIO DE COLOMBIA

- Según Artes (2002)

- Muestreo repetido.
- Muestreo panel.
- Muestreo en ocasiones sucesivas. (Cochran 1977)

- Características

- Dos muestras diferentes de la misma población en dos instantes de tiempo distintos  $t_1 < t_2$ .
- Medir la evolución de la característica poblacional de interés a lo largo del tiempo.
- Dos mediciones (i.e: variables  $x$  e  $y$ ) diferentes, pero dependientes ya que se realizan en la misma unidad muestral.
- Jessen (1942) fue el primero en considerar en un estudio real la aplicación.

- Ventajas

- Reduce los costos.
- Aumenta la precisión de los estimadores.
- Desgaste de la muestra.



# Notación: Muestreo en dos fases

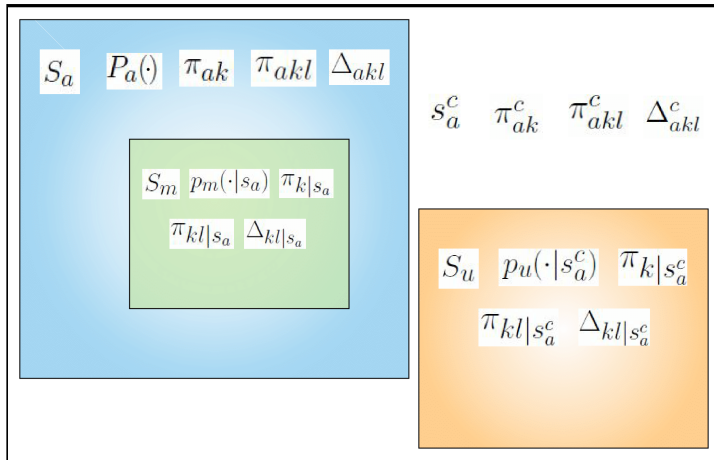


Figura 1: Notación



# Parámetros de interés

Särndal (1992, Pág. 369) propone algunos parámetros de interés de la forma:

$$t = \phi t_z + \psi t_y,$$

donde  $t_z = \sum_U z_k$  ,  $t_y = \sum_U y_k$ ,  $\phi$  y  $\psi$  son constantes.

- $\phi = 0, \psi = 1$ , entonces  $t = t_y$ .
- $\phi = -1, \psi = 1$ , entonces  $t = t_y - t_z$ .
- $\phi = -1/t_z, \psi = 1/t_z$ , entonces  $t = (t_y - t_z)/t_z$ .
- $\phi = 1, \psi = 1$ , entonces  $t = t_y + t_z$ .



# Técnica de conteo de ítem propuesta por Hussain, Shah & Shabbir (2012) (TCIH)

- Hussain & Shabbir (2010).
- No requiere de dos submuestras de tamaños  $n_1$  y  $n_2$ .
- $J$  items, con  $J \geq 2$ .
- Dos preguntas para el  $j$ -ésimo ítem:
  - Pregunta no sensible ( $T_j$ )
  - Pregunta con característica sensible ( $S$ )
- Contar 1 si posee alguna de las características  $T_j$  o  $S$  y 0 en otro caso e informar el conteo.



# Técnica de conteo de ítem propuesta por Hussain et al. (2012) (TCIH)

- $Z_i$ : Conteo total del  $i$ -ésimo encuestado

$$Z_i = \sum_{j=1}^J \alpha_j,$$

- $\alpha_j$  puede tomar valores 1 o 0 con probabilidades  $(\pi + \theta_j - \pi\theta_j)$  y  $(1 - \pi - \theta_j + \pi\theta_j)$ , respectivamente.

$$E(Z_i) = \sum_{j=1}^J E(\alpha_j)$$

$$= J\pi + \sum_{j=1}^J \theta_j - \pi \sum_{j=1}^J \theta_j$$

$$= \left( J - \sum_{j=1}^J \theta_j \right) \pi + \sum_{j=1}^J \theta_j.$$



# Técnica de conteo de ítem propuesta por Hussain et al. (2012) (TCIH)

- $Z_i$ : Conteo total del  $i$ -ésimo encuestado

$$Z_i = \sum_{j=1}^J \alpha_j,$$

- $\alpha_j$  puede tomar valores 1 o 0 con probabilidades  $(\pi + \theta_j - \pi\theta_j)$  y  $(1 - \pi - \theta_j + \pi\theta_j)$ , respectivamente.

$$E(Z_i) = \sum_{j=1}^J E(\alpha_j)$$

$$= J\pi + \sum_{j=1}^J \theta_j - \pi \sum_{j=1}^J \theta_j$$

$$= \left( J - \sum_{j=1}^J \theta_j \right) \pi + \sum_{j=1}^J \theta_j.$$



# Técnica de conteo de ítem propuesta por Hussain et al. (2012) (TCIH)

Hussain & Shabbir (2010) sugieren un estimador insesgado

$$\hat{\pi}_p = \frac{\bar{Z} - \sum_{j=1}^J \theta_j}{J - \sum_{j=1}^J \theta_j}.$$

$$V(\hat{\pi}_p) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)}{n \left( J - \sum_{j=1}^J \theta_j \right)^2}$$

$$\left\{ \left( \sum_{j=1}^J \theta_j \right) \left( 1 - \sum_{j=1}^J \theta_j \right) + \sum_{\substack{j,k=1 \\ j \neq k}}^J \theta_j \theta_k \right\}.$$

## Recomendaciones y resultados

- Sensibilización previa.
- Explicación de la metodología utilizada.
- No presencia del entrevistador.
- Mejores resultados que la TCI tradicional y la TRA.



UNIVERSIDAD SANTO TOMÁS  
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

# Técnica de conteo de ítem propuesta por Hussain et al. (2012) (TCIH)

Hussain & Shabbir (2010) sugieren un estimador insesgado

$$\hat{\pi}_p = \frac{\bar{Z} - \sum_{j=1}^J \theta_j}{J - \sum_{j=1}^J \theta_j}.$$

$$V(\hat{\pi}_p) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)}{n \left( J - \sum_{j=1}^J \theta_j \right)^2}$$

$$\left\{ \left( \sum_{j=1}^J \theta_j \right) \left( 1 - \sum_{j=1}^J \theta_j \right) + \sum_{\substack{j,k=1 \\ j \neq k}}^J \theta_j \theta_k \right\}.$$

## Recomendaciones y resultados

- Sensibilización previa.
- Explicación de la metodología utilizada.
- No presencia del entrevistador.
- Mejores resultados que la TCI tradicional y la TRA.

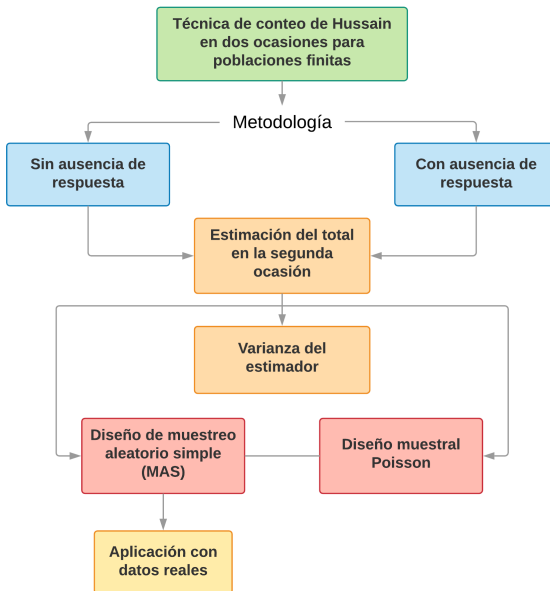


# Contenido

- 1 Introducción
- 2 Marco Teórico
- 3 Método de estimación
- 4 Sumarización
- 5 Conclusiones



UNIVERSIDAD SANTO TOMÁS  
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA



# Sin ausencia de respuesta



UNIVERSIDAD SANTO TOMÁS  
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

# Notación

- Se considera un muestreo en dos ocasiones de una población finita  $U = \{1, \dots, k, \dots, N\}$
- La variable de estudio será denotada por  $Z$  en la primera ocasión y por  $Y$  para la segunda ocasión.

$$\alpha_{1kj} = \begin{cases} 1 & \text{Si el individuo } k \text{ posee alguna de las dos características en el ítem } j \text{ de la ocasión 1.} \\ 0 & \text{Si el individuo } k \text{ no posee ninguna de las dos características en el ítem } j \text{ de la ocasión 1.} \end{cases}$$

$$\alpha_{2kg} = \begin{cases} 1 & \text{Si el individuo } k \text{ posee alguna de las dos características en el ítem } g \text{ de la ocasión 2.} \\ 0 & \text{Si el individuo } k \text{ no posee ninguna de las dos características en el ítem } g \text{ de la ocasión 2.} \end{cases}$$

# Metodología sin ausencia de respuesta

En la primera ocasión, una muestra  $S_a$  es seleccionada usando el diseño muestral  $p_a(\cdot)$  y la variable  $Z$  es medida para todos los elementos en  $S_a$ . Las probabilidades de inclusión de primer y segundo orden asociadas al diseño muestral  $p_a(\cdot)$  son denotadas por  $\pi_{ak}$  y  $\pi_{akl}$  respectivamente, y  $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$ .

Soberanis, Trujillo & González (2013)

$$\hat{t}_Z = \frac{1}{J - \sum_{j=1}^J \theta_j} \left( \sum_{s_a} \frac{Z_k}{\pi_{ak}} - N \sum_{j=1}^J \theta_j \right) \quad (1)$$

Con  $Z_k = \sum_{j=1}^J \alpha_{1kj}$

- $\theta_j$ : Proporción de individuos con la característica no sensible  $j$ .
- $J$ : Número de ítems presentado en el formulario.

# Estimador propuesto

$$\hat{t}_1 = \hat{t}_{KS_a} + \hat{t}_{DS_m}, \quad (2)$$

donde

$$\hat{t}_{KS_a} = B_1 \left( \sum_{S_a} \frac{Z_k}{\pi_{ak}} - N \sum_{j=1}^J \theta_j \right)$$

$$\hat{t}_{DS_m} = \left[ \sum_{S_m} \frac{D_k}{\pi_{ak} \pi_{k/S_a}} - N \left( B_2 \sum_{g=1}^G \lambda_g - B_1 \sum_{j=1}^J \theta_j \right) \right]$$

con

$$D_k = B_2 Y_k - B_1 Z_k, \quad B_1 = \frac{K}{J - \sum_{j=1}^J \theta_j} \quad \text{y} \quad B_2 = \frac{1}{G - \sum_{g=1}^G \lambda_g}.$$

## Estimador propuesto

Usando la información de la primera ocasión y de la muestra traslapada de la segunda ocasión,  $\hat{t}_1$  se puede reescribir como:

$$\hat{t}_1 = B_1 \sum_{S_a} \frac{Z_k}{\pi_{ak}} + \sum_{S_m} \frac{D_k}{\pi_{ak} \pi_{k/S_a}} - NB_2 \sum_{g=1}^G \lambda_g.$$

Usando la información de la muestra no traslapada, recolectada en la segunda ocasión

$$\hat{t}_2 = \hat{t}_{Y S_u} = B_2 \left( \sum_{S_u} \frac{Y_k}{\pi_{ak}^c \pi_{k/S_a}^c} - N \sum_{g=1}^G \lambda_g \right).$$

Finalmente, se propone (3) como un estimador compuesto para estimar  $t_Y$ .

$$\hat{t}_Y = w_1 \hat{t}_1 + w_2 \hat{t}_2, \quad (3)$$

con  $w_1$  y  $w_2$  constantes no negativas tal que  $w_1 + w_2 = 1$ .

# Estimador propuesto

## Proposición

*La covarianza entre  $Z_k$  y  $Y_k$ , bajo la TCIH es:*

$$Cov_{TCIH}(Y_k, Z_k) = \pi_1(\pi_{1|1} - \pi_2) \sum_{g=1}^G \sum_{j=1}^J (1 - \lambda_g)(1 - \theta_j). \quad (4)$$

## Proposición

*El valor esperado del estimador compuesto (3) está dado por:*

$$E(\hat{t}_Y) = E(w_1 \hat{t}_1 + w_2 \hat{t}_2) = N\pi_2. \quad (5)$$



# Varianza del estimador $\hat{t}_Y$

## Proposición

La varianza del estimador compuesto (3) es:

$$V(\hat{t}_Y) = w_1^2 V(\hat{t}_1) + w_2^2 V(\hat{t}_2) + 2w_1w_2Cov(\hat{t}_1, \hat{t}_2), \quad (6)$$

donde

$$\begin{aligned} V(\hat{t}_1) = & B_2^2 E_{TCIH}^2(Y_k) \left( \sum_U \sum_U \frac{\Delta_{akl}}{\pi_{ak} \pi_{al}} \right) \\ & + E_{TCIH}^2(D_k) E_{p_a} \left( \sum_{S_a} \sum_{S_a} \frac{\Delta_{kl/S_a}}{\pi_{ak} \pi_{k/S_a} \pi_{al} \pi_{l/S_a}} \right) \\ & + B_1^2 V_{TCIH}(Z_k) \sum_U \frac{1}{\pi_{ak}} + V_{TCIH}(D_k) E_{p_a} E_{p_m} \left( \sum_{S_m} \frac{1}{\pi_{ak}^2 \pi_{k/S_a}^2} \right) \end{aligned}$$

# Varianza del estimador $\hat{t}_Y$

$$\begin{aligned}
 V(\hat{t}_2) = & B_2^2 E_{TCIH}^2(Y_k) \left[ \sum_U \sum_U \frac{\Delta_{akl}^c}{\pi_{ak}^c \pi_{al}^c} \right. \\
 & + E_{TCIH}^2(Y_k) E_{p_a} \left( \sum_{S_a^c} \sum_{S_a^c} \frac{\Delta_{kl}/S_a^c}{\pi_{ak}^c \pi_{k/S_a^c} \pi_{al}^c \pi_{l/S_a^c}} \right) \Big] \\
 & + B_2^2 V_{TCIH}(Y_k) E_{p_a} E_{p_u} \left( \sum_{S_u} \frac{1}{(\pi_{ak}^c \pi_{k/S_a^c})^2} \right)
 \end{aligned}$$

$$Cov(\hat{t}_1, \hat{t}_2) = -B_2^2 E_{TCIH}^2(Y_k) \left( \sum_U \sum_U \frac{\Delta_{akl}}{\pi_{ak} \pi_{al}^c} \right)$$

# Valor de $K$ y $w_1$ que minimizan la varianza dada en (6)

## Proposición

El valor de  $K$  que minimiza la varianza del estimador (3) está dado por:

$$K_{opt} = \frac{B_2 \left( J - \sum_{j=1}^J \theta_j \right) A}{B} \quad (7)$$

donde,

$$\begin{aligned} A = & Cov_{TCIH}(Y_k, Z_k) E_{pa} E_{pm} \left( \sum_{S_m} \frac{1}{\pi_{ak}^2 \pi_{k/S_a}^2} \right) \\ & + E_{TCIH}(Z_k) E_{TCIH}(Y_k) E_{pa} \left( \sum_{S_a} \sum_{S_a} \frac{\Delta_{kl/S_a}}{\pi_{ak} \pi_{k/S_a} \pi_{al} \pi_{l/S_a}} \right) \end{aligned}$$

$$B = V_{TCIH}(Z_k) \sum_U \frac{1}{\pi_{ak}} + E_{TCIH}^2(Y_k) E_{p_a} \left( \sum_{S_a} \sum_{S_a} \frac{\Delta_{kl/S_a}}{\pi_{ak} \pi_{k/S_a} \pi_{al} \pi_{l/S_a}} \right) \\ + V_{TCIH}(Z_k) E_{p_a} E_{p_m} \left( \sum_{S_m} \frac{1}{\pi_{ak}^2 \pi_{k/S_a}^2} \right)$$

Valor  $w_1$ , que minimiza la varianza en (6)

$$w_1 = \frac{V(\hat{t}_2) - Cov(\hat{t}_1, \hat{t}_2)}{V(\hat{t}_1) + V(\hat{t}_2) - 2Cov(\hat{t}_1, \hat{t}_2)} \quad (8)$$



$$B = V_{TCIH}(Z_k) \sum_U \frac{1}{\pi_{ak}} + E_{TCIH}^2(Y_k) E_{p_a} \left( \sum_{S_a} \sum_{S_a} \frac{\Delta_{kl/S_a}}{\pi_{ak} \pi_{k/S_a} \pi_{al} \pi_{l/S_a}} \right) \\ + V_{TCIH}(Z_k) E_{p_a} E_{p_m} \left( \sum_{S_m} \frac{1}{\pi_{ak}^2 \pi_{k/S_a}^2} \right)$$

Valor  $w_1$ , que minimiza la varianza en (6)

$$w_1 = \frac{V(\hat{t}_2) - Cov(\hat{t}_1, \hat{t}_2)}{V(\hat{t}_1) + V(\hat{t}_2) - 2Cov(\hat{t}_1, \hat{t}_2)} \quad (8)$$



# Contenido

- 1 Introducción
- 2 Marco Teórico
- 3 Método de estimación
- 4 Simulación**
- 5 Conclusiones



UNIVERSIDAD SANTO TOMÁS  
PRIMER CUAESTRO UNIVERSITARIO DE COLOMBIA

# Escenario de simulación

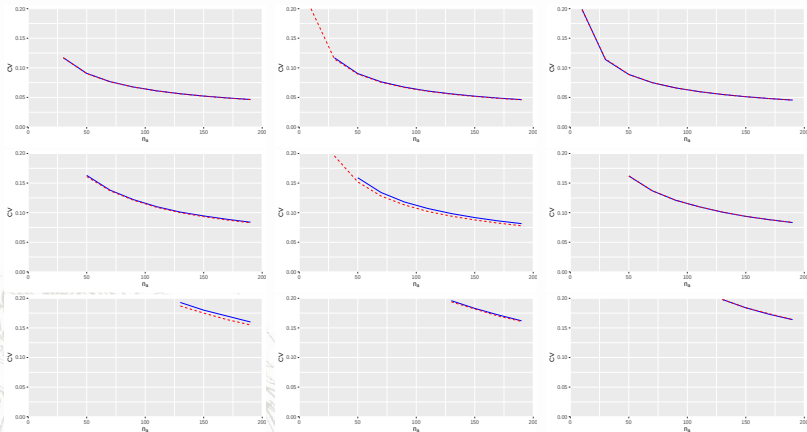
Cuadro 1: Escenario de simulación

|      |         |         |     |      |
|------|---------|---------|-----|------|
| 0.75 | $\pi_2$ |         |     |      |
| 0.5  |         |         |     |      |
| 0.25 |         |         |     |      |
|      |         | $\pi_1$ |     |      |
|      |         | 0.25    | 0.5 | 0.75 |

- Los  $n_a$  alternan en una secuencia de 20, comenzando en 10 y terminando en 190.
- El tamaño de la muestra  $n_u$  es igual a  $n_a$ .
- Los porcentajes de traslape que se utilizaron fueron de  $\Gamma = 20\%$  y  $\Gamma = 80\%$ .

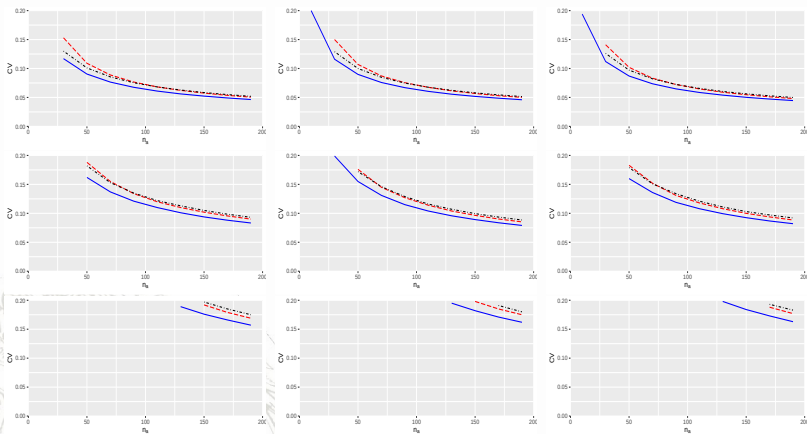


# Resultados diseño MAS



**Figura 2:** Relación entre el tamaño de muestra ( $n_a$ ) y el coeficiente de variación de  $\hat{t}_Y$  ( $CV$ ) para un diseño MAS.  $\Gamma = 0.2$  (línea sólida) y  $\Gamma = 0.8$  (línea punteada).

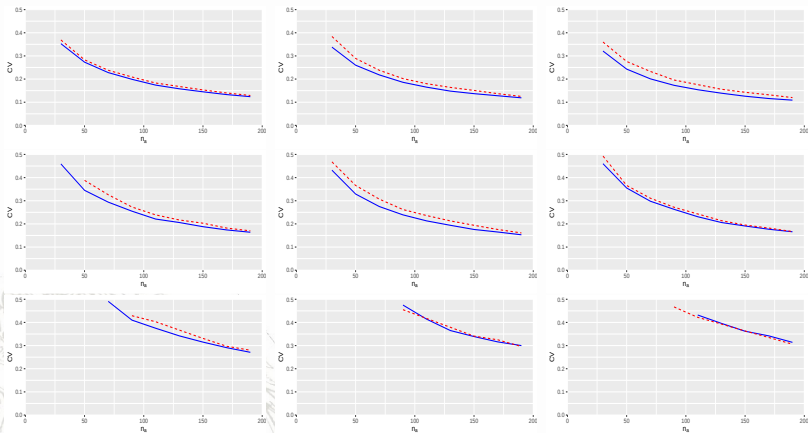
# Diseño MAS con ausencia de respuesta



**Figura 3:** Relación entre el tamaño de muestra ( $n_a$ ) y el coeficiente de variación de  $\hat{t}_Y$  ( $CV$ ) para un diseño MAS con ausencia de respuesta.

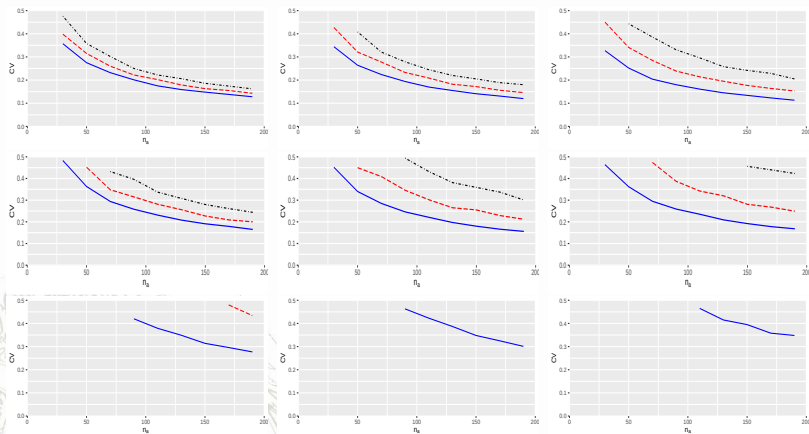
—  $\%NR = 0\%$ , - - -  $\%NR = 10\%$ , · · ·  $\%NR = 20\%$ .

# Resultados diseño Poisson



**Figura 4:** Relación entre el tamaño de muestra ( $n_a$ ) y el coeficiente de variación de  $\hat{t}_Y$  ( $CV$ ) para un diseño Poisson.  $\Gamma = 0.2$  (línea sólida) y  $\Gamma = 0.8$  (línea punteada).

# Diseño Poisson con ausencia de respuesta



**Figura 5:** Relación entre el tamaño de muestra ( $n_a$ ) y el coeficiente de variación(CV) de  $\hat{t}_Y$  (CV) para un diseño Poisson con ausencia de respuesta.

—  $\%NR = 0\%$ , - - -  $\%NR = 10\%$ , - · - ·  $\%NR = 20\%$ .

# Contenido

- 1 Introducción
- 2 Marco Teórico
- 3 Método de estimación
- 4 Sumarización
- 5 Conclusiones



UNIVERSIDAD SANTO TOMÁS  
PRIMER CUAESTRO UNIVERSITARIO DE COLOMBIA

# Conclusiones

- Propuesta de estimación con su respectiva varianza teórica para el total de individuos con una característica sensible.
- Vía simulación, se logró ilustrar el comportamiento del método de estimación propuesto y algunas propiedades de los estimadores.
- A falta de un estimador para la varianza de los estimadores propuestos en este trabajo, se recomienda utilizar el método de estimación Bootstrap, tal como se utilizó en el ejercicio de aplicación.
- Con el fin de que los encuestados comprendan perfectamente el método propuesto, se aconseja realizar una motivación previa haciendo énfasis en dos aspectos:
  - 1 El encuestado debe contestar **cuántas** de las preguntas son verdaderas mas no cuales.
  - 2 Explicar claramente la definición de “o” como conector lógico en una proposición.

# References I

-  Artes, E. y Garcia, A. (2002), 'Aportaciones al muestreo sucesivo', *Revista Colombiana de Estadística* **25 No 2**, 145 – 158.
-  Cochran, W. G. (1977), 'Sampling techniques-3'.
-  Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W. & Ezzati, T. M. (1991), *The item count technique as a method of indirect questioning: A review of its development and a case study application*, Wiley Online Library, pp. 185–210.
-  Hussain, Z. & Shabbir, J. (2010), 'On item count technique in survey sampling', *Journal of Informatics and Mathematical Sciences* **2(2 & 3)**, 161–169.
-  Hussain, Z., Shah, E. A. & Shabbir, J. (2012), 'An Alternative Item Count Technique in Sensitive Surveys', *Revista Colombiana de Estadística* **35(1)**, 39–54.

## References II

- Jessen, R. J. (1942), 'Statistical investigation of a sample survey for obtaining farm facts\ raymond j. jessen'.
- Karlan, D. S. & Zinman, J. (2012), 'List randomization for sensitive behavior: An application for measuring use of loan proceeds', *Journal of Development Economics* **98**(1), 71–75.
- Kuha, J. & Jackson, J. (2014), 'The item count method for sensitive survey questions: modelling criminal behaviour', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **63**(2), 321–341.
- Miller, J. D. (1984), A New Survey Technique for Studying Deviant Behavior, PhD thesis, The George Washington University Miller.
- Särndal, Carl; Swensson, B. W. J. (1992), *Model Assisted Survey Sampling*.



## References III

-  Sliderman, P., Tetlock, P. & Piazza, T. (1992), 'Codebook for the 1991 national race and politics survey', *Survey Research Center, Berkeley, CA, available at <http://sda.berkeley.edu>* **D 3**.
-  Siberanis, V., Trujillo, L. & González, L. (2013), Item count techniques in sampling from finite population, Technical report.
-  Warner, S. L. (1965), 'Randomized response: A survey technique for eliminating evasive answer bias', *Journal of the American Statistical Association* **60**(309), 63–69.

