Una técnica de conteo de items como alternativa a las respuestas aleatorizadas en muestreo de poblaciones finitas sin reemplazo

Víctor H. Soberanis Cruz*
Leonardo Trujillo**
Luz Mery González***

Resumen

Las Técnicas de Conteo de Items (ICT) han sido desarrolladas, al igual que las Técnicas de Respuestas Aleatorizadas (RRT), para estudiar asuntos sensitivos. El objetivo de este trabajo es implementar la Técnica de Conteo de Items de Zawar Hussain, Ejaz Ali Shah y Javid Shabbir (2012) en muestreo de poblaciones finitas sin reemplazo, para estimar el total de individuos en la población que poseen una cierta característica sensible A. Esta técnica tiene la ventaja, entre otros puntos, de que no requiere dos submuestras como en la Técnica de Conteo de Items de Kosuke Imai (2010). Otra ventaja, es que es mucho más fácil de aplicar que las técnicas de Respuestas Aleatorizadas. En el entendido de que el uso de variables auxiliares hace más eficientes las técnicas de muestreo en general, en este trabajo no nos ocuparemos de la implementación de tales variables, toda vez que pretendemos la aplicación ligera de una ICT. Compararemos entonces a Hussain *et al.*, (HSSICT) con la técnica MU de Respuestas Aleatorizadas (Soberanis *et al.*, 2008) que es más eficiente que la RRT de Warner (1965).

*Departamento de Ciencias, División de Ciencias e Ingeniería, Universidad de Quintana Roo, México.
**Departamento de Estadística. Universidad Nacional de Colombia.
***Departamento de Estadística. Universidad Nacional de Colombia.

Abstract

Item Count (ICT) and Randomized Response (RR) techniques have been developed in the literature in order to estimate parameters related to sensitive questions. The aim of this paper is to implement one available

item count technique (Hussain, *et al.*, 2012) in a finite population sampling without replacement set up, in order to estimate the total of individuals in the population having a particular sensitive characteristic. The main diference between Imai's (2010) ICT and Hussain's ICT is that the former requires two unmatched samples whereas the later requires one sample being taken from the population. Although Hussain's ICT is easer to apply than Morton-Un related RR technique (Soberanis *et al.*, 2008), we compare the efficiency of these two approaches. Estimators for the population total as their corresponding variance -estimators are obtained for a general finite population sampling design without replacement set up.

Introducción

Cuando se trata de obtener información sensitiva, es común el problema de la no respuesta o que la veracidad de la misma esté en duda. Una ingeniosa alternativa a las preguntas directas fue introducida por Warner (1965), conocida como Técnica de Respuestas Aleatorizadas. Para una buena revisión respecto al desarrollo de las RRT se puede consultar Tracy & Mangat (1996) y Chaudhuri & Mukherjee (1988). Sin embargo, Geurts (1980) ha reportado que las RRT requieren de mayor tamaño de muestra para obtener intervalos de confianza comparables con los obtenidos por muestreo directo. Las experiencias revelan que en necesario más tiempo para administrar y explicar la RRT a los encuestados. Dalton & Metzger (1992) reportan que las RRT pueden no ser efectivas en encuestas por teléfono o correo electrónico. También Casper & Lessler (1989) dicen que la aceptación de la RRT por parte del entrevistado más de las veces es complicado. Las ICT (Droitcour, *et al.*, 1991) son una alternativa a las RRT.

Palabrasclave: diferencia diseño-base,población finita,técnicas de conteo de items, técnicas de respuestas aleatorizadas, preguntas delicadas, muestreo de la encuesta.

1. La técnica de conteo de ítems de Hussain

Consideremos una población consistente de N elementos y para simplificar, el k-ésimo elemento de la población será representado por su etiqueta $k\ k$. De manera que denotaremos a la población finita como:

$$U = \{1,2, \dots, N\}$$

En este trabajo el tamaño de la población $N\ N$ se supondrá conocido. Sea $y\ y$ una variable que mide alguna característica sensitiva, y sea $y_k\, y_k$ el valor de $y\ y$ para el k-ésimo elemento de la población. Así $y_k\, y_k$ es desconocida pero no aleatoria. Además $y_k = 1\ y_k = 1$ si el k-ésimo individuo de la población tiene la característica sensitiva A, y $y_k = 0\ \ y_k =$ $t_A = \sum_U y_k$ simo individuo no tiene la característica sensitiva A. Lo que se desea es estimar $t_A = \sum_U y_k$, total de los individuos en la población con la característica sensitiva A. La selección de la muestra es mediante el diseño $p(s)\ p(s)$ con probabilidades positivas de inclusión $\pi_k\ \pi_k$ y $\pi_{kl}\, \pi_{kl}$ donde:

$$\pi_k = \Pr\{S \ni k\} = \sum_{S \ni K} p(s) \quad y \quad \pi_{kl} = \Pr(S \ni k \& l) = \sum_{s \ni k \& l} p(s)$$

La ICT que proponen Hussain *et al.*, es como sigue: a cada elemento $k\ k$ en la muestra $S\ S$ se le proporciona un cuestionario con g ítems. El $j\ j$ -ésimo ítems consta de dos preguntas, una pregunta $F_j\,F_j$ la cual es inocua y la pregunta sensitiva A. Al entrevistado se le pide que le asigne el valor uno al ítem $j\ j$ si éste se identifica con al menos una de las preguntas del ítem, le asignará el valor cero de otro modo. El entrevistado reportará al entrevistador su conteo total basado en el cuestionario completo.

En el modelo de Hussain se tiene, para cada $k\ k$ en la muestra $S\ S$, una lista (Cuestionario) con una configuración como la que se muestra en la siguiente Tabla:

1	F_1	A	α_{k1}
2	F_2	A	α_{k2}
⋮	⋮	⋮	⋮
J	F_j	A	α_{kj}
⋮	⋮	⋮	⋮
g	F_g	A	α_{kg}

Donde las $F_j\,F_j$ son eventos que denotan las características innocuas y A es el evento que denota a la característica sensible, además:

$$a_{kj} = \begin{cases} 1 \;\; si \;\; k \in F_j \cup A; \\ 0 \;\; de\ otro\ modo \end{cases}$$

De modo que

$$\mathbf{Pr}\{a_{kj} = 1\} = \theta_j + \pi - \theta_j\pi = \pi + (1-\pi)\theta_j$$

Donde: $\pi = \frac{t_A}{N}\pi = \frac{t_A}{N}$ y $\theta_j = \mathbf{Pr}\{k \in F_j\}\theta_j = \mathbf{Pr}\{k \in F_j\}$. En este trabajo asumiremos que las $\theta_j\,\theta_j$ son conocidas.

Sea:

$$Z_k = \sum_{j=1}^{g} \alpha_{kl}$$

El estimador-$\pi\ \pi$ de Hussain para el total de individuos con la característica sensible viene dado por:

$$\hat{t}_{A,H,\pi} = \frac{\mathbf{1}}{g - \theta.} \sum_{S} \frac{Z_k - \theta.}{\pi_k}$$

$$\theta. = \sum_{j=1}^{g} \theta_j \; \theta. = \sum_{j=1}^{g} \theta_j$$

Donde:

Este estimulador $\hat{t}_{A,H,\pi}$ $\hat{t}_{A,H,\pi}$ es insesgado para $\hat{t}_A$ $\hat{t}_A$:

$$E\left(\hat{t}_{A,H,\pi}\right) = E_p E_{HSS}\left(\hat{t}_{A,H,\pi}\right)$$

$$= \frac{\mathbf{1}}{g - \theta.} E_p E_{HSS}\left[\sum_{U} I_k(S) \frac{Z_k - \theta.}{\pi_k}\right]$$

$$= \frac{\mathbf{1}}{g - \theta.} E_p \left[\sum_{U} I_k(S) \frac{E_{HSS}(Z_k) - \theta.}{\pi_k}\right]$$

$$= \frac{t_A}{N} E_p \left[\sum_{U} I_k(S) \frac{\mathbf{1}}{\pi_k}\right]$$

$$= t_A$$

de los estimadores-π π nos permite obtener tanto la varianza del estimador $\hat{t}_{A,H,\pi}$ $\hat{t}_{A,H,\pi}$ como un estimador de la varianza del estimador $\hat{t}_{A,H,\pi}$ $\hat{t}_{A,H,\pi}$.

Varianza del estimador $\hat{\boldsymbol{t}}_{\boldsymbol{A,H,\pi}}$ $\hat{\boldsymbol{t}}_{\boldsymbol{A,H,\pi}}$

$$V\left(\hat{t}_{A,H,\pi}\right) = E_p V_{HSS}\left(\hat{t}_{A,H,\pi}\right) + V_p E_{HSS}\left(\hat{t}_{A,H,\pi}\right)$$

Tenemos:

$$E_{HSS}\left(\alpha_{kj}\alpha'_{kj}\right) = \Pr\left\{k \in A \cup \left(F_j \cap F'_j\right)\right\}$$
$$= \Pr\{k \in A\} + \Pr\left\{k \in \left(F_j \cap F'_j\right)\right\} - \Pr\left\{k \in A \cap \left(F_j \cap F'_j\right)\right\}$$
$$= \pi + \theta_j\theta'_j - \pi\theta_j\theta_{j'} = \pi + (1 - \pi)\theta_j\theta'_j$$

De manera que:

$$E_{HSS}(Z_k^2) = E_{HSS}\left(\sum_{j=1}^{g}\alpha_{kj}^2 + \sum_{j\neq j'}\sum \alpha_{kj}\alpha'_{kj}\right)$$

$$= E_{HSS}\left(\sum_{j=1}^{g}\alpha_{kj} + \sum_{j\neq j'}\sum \alpha_{kj}\alpha'_{kj}\right)$$

$$= \sum_{j=1}^{g}[\pi + (1-\pi)\theta_j] + \sum_{j\neq j'}\sum\left[\pi + (1-\pi)\theta_j\theta_{j'}\right]$$

$$= \theta_. + (g-\theta_.)\pi + g(1-g)\pi + (1-\pi)\sum\sum_{j\neq j'}\theta_j\theta'_j$$

Y:

$$V_{HSS}(Z_k) = E_{HSS}(Z_k^2) - [E_{HSS}(Z_k)]^2$$

$$= \theta_. + (g-\theta_.)\pi + g(1-g)\pi + (1-\pi)\sum_{j\neq j'}\theta_j\theta'_j - [g\pi + (1-\pi)\theta_.]^2$$

$$= (1-\pi)\theta_. + g\pi_. + g(1-g)\pi + (1-\pi)\sum_{j\neq j'}\theta_j\theta'_j - [g\pi + (1-\pi)\theta_.]^2$$

$$\equiv V_{0H}(\pi)(\textbf{Independiente de } k)$$

Por tanto:

$$E_p V_{HSS}(\hat{t}_{A,H,\pi}) = E_p V_{HSS}\left(\sum_S \frac{z_k - \theta_.}{(g-\theta_.)\pi_k}\right)$$

$$= E_p\left[\sum_U \frac{I_k(S)}{(g-\theta_.)^2\pi_k^2} V_{HSS}(Z_k)\right]$$

$$= \frac{V_{0H}}{(g-\theta_.)^2}\sum_U \frac{1}{\pi_k}$$

También tenemos que:

$$E_{HSS}(\hat{t}_{A,H,\pi}) = E_{HSS}\left(\sum_s \frac{Z_k - \theta.}{(g-\theta.)\pi_k}\right)$$

$$= \left[\sum_U \frac{I_k(S)}{(g-\theta.)\pi_k} E_{HSS}(Z_k - \theta.)\right]$$

$$= \left[\sum_U \frac{I_k(S)}{(g-\theta.)\pi_k} E_{HSS}(Z_k - \theta.)\right]$$

$$= \sum_U \frac{I_k(S)}{\pi_k}\pi = \sum_U \frac{I_k(S)}{\pi_k}\pi$$

$$V_p E_{HSS}(\hat{t}_{A,H,\pi}) = V_p\left[\sum_U \frac{I_k(S)}{\pi_k}\pi\right]$$

$$= \left(\frac{t_A}{N}\right)^{2\Sigma\Sigma_U \Delta_{kl}\frac{1}{\pi_k}\frac{1}{\pi_l}}$$

$$= \left(\frac{t_A}{N}\right)^{2\Sigma\Sigma_U \Delta_{kl}\frac{1}{\pi_k}\frac{1}{\pi_l}}$$

Así que:

$$V(\hat{t}_{A,H,\pi}) = \frac{V_{0H}}{(g-\theta.)^2}\sum_U \frac{1}{\pi_k} + \left(\frac{t_A}{N}\right)^2 \sum\sum_U \Delta_{kl}\frac{1}{\pi_k}\frac{1}{\pi_l}$$

Un estimador de varianza podría ser:

$$\hat{V}(\hat{t}_{A,H,\pi}) = \frac{\hat{V}_{0H}}{(g-\theta.)^2}\sum_U \frac{1}{\pi_k} + \left(\frac{\hat{t}_{A,H,\pi}}{N}\right)^2 \sum\sum_U \Delta_{kl}\frac{1}{\pi_k}\frac{1}{\pi_l}$$

Donde: $\hat{V}_{OH} = V_{0H}(\hat{t}_{A,H,\pi})\hat{V}_{OH} = V_{0H}(\hat{t}_{A,H,\pi})$.

2. Simulación

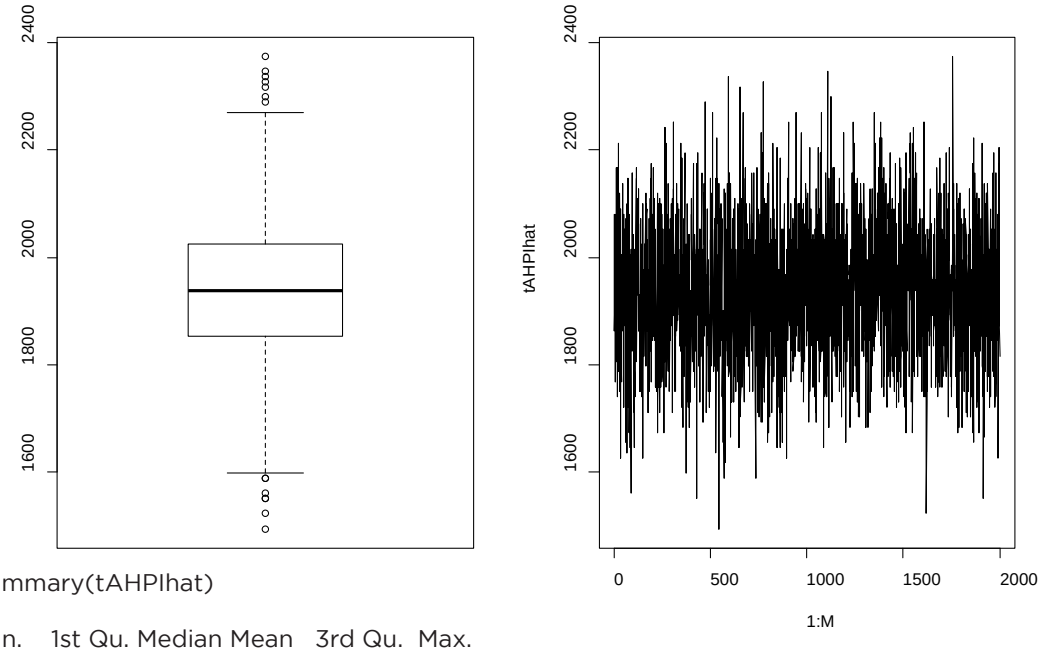
En esta sección mostramos vía simulación las bondades del estimador que proponemos para el total de individuos en la población con la característica sensible A. Los códigos están en R y parte de estos pueden ser usados en una aplicación real y mostrar los resultados obtenidos con base en la muestra.

#Técnica de conteo de items de Hussain

```
N<-7000

A<-2000

g<-5

theta1<-0.30

theta2<-0.40

theta3<-0.15

theta4<-0.20

theta5<-0.25

thetaVec<-c(theta1,theta2,theta3,theta4,theta5)

F1<-sample(0:1,N,replace=T,prob=c(1-theta1,theta1))

F2<-sample(0:1,N,replace=T,prob=c(1-theta2,theta2))

F3<-sample(0:1,N,replace=T,prob=c(1-theta3,theta3))

F4<-sample(0:1,N,replace=T,prob=c(1-theta4,theta4))

F5<-sample(0:1,N,replace=T,prob=c(1-theta5,theta5))

thetaDot <-sum(thetaVec)

U<-c(rep(0,N-A),rep(1,A))

U<-sample(U,N)

dfFA<-data.frame(F1,F2,F3,F4,F5,U)

ZU<-rep(-1,N)

for (i in 1:N) if (dfFA(i,6)<-1) (ZU(i)<-sum(dfFA(i,) )) else (ZU(i)<-5)

n<-200

M <- 2000

tAHPIhat<-rep(0,M)

for (j in 1:M) {

                ZS<-sample(ZU,n,replace=FALSE)

                ZS_thetaD<-(N□n)*(ZS-thetaDot)
```

```
        tAHPIhat(j) <-(1/(g-thetaDot))*sum(ZS_thetaD)
        }
summary(tAHPIhat)
sqrt(var(tAHPIhat))
par(mfrow=c(1,2))
boxplot(tAHPIhat)
plot(1:M,tAHPIhat,type= "l")
```

2.1 Resultados de la simulación



summary(tAHPIhat)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1239 1864 2024 2028 2195 2923

sqrt(var(tAHPIhat))=244.46

3. Modelo Mu de respuestas aleatorizadas

3.1 El procedimiento de muestreo

Queremos estimar $t_A = \sum_U y_k \, t_A = \sum_U y_k$ considerando que $t_W = \sum_U w_k \, t_W = \sum_U w_k$, total de la característica no sensitiva W no relacionada con la característica sensitiva A, es conocido.

El procedimiento de muestreo es como sigue:

(a) Se extrae una muestra de tamaño n de acuerdo mal diseño de muestreo $p(s)$ $p(s)$

(b) La metodología de Preguntas no relacionadas (modificada) es: El mEcanismo aleatorio RC será tal que $Q_A\,Q_A$ es elegido con probabilidad P y $Q_W\,Q_W$ con probabilidad $1-P\;1-P$. La modificación consiste en que tanto $Q_A\,Q_A$ como $Q_W\,Q_W$ vienen acompañadas de la pregunta ¿ Pertenece usted al grupo G1 o al grupo G2?, donde G1=$(A\cap W\;\;(A\cap W\;)\cup(\overline{A}\cap\overline{W})\cup(\overline{A}\cap\overline{W})$ y G2= $(A\cap\overline{W})\cup(\overline{A}\cap W).(A\cap\overline{W})\cup(\overline{A}\cap W).$

Definimos

$$d_k=\begin{cases}0\;\; si\;\; k\in G1\Leftrightarrow y_k=w_k\\1\;\; si\;\; k\in G2\Leftrightarrow y_k\neq w_k\end{cases}$$

Nótese que $d_k\,d_k$ es independiente del mecanismo aleatorio MC; es decir, $E_{RC}(d_k)=d_k$ $E_{RC}(d_k)=d_k$. Recordemos también que la característica W es no sensitiva, es innocua.

3.2 El estimador y la varianza del estimador

Sea

$$Z_k=\begin{cases}y_k\;\; con\;\; probabilidad & P\\w_k\;\; con\; probabilidad & 1-P\end{cases}$$

Así

$$E_{RC}(Z_k)=y_kP+w_k(1-P)$$

$$\equiv\theta_k\equiv\theta_k$$

Si $t_\theta=\sum_U\theta_k\,t_\theta=\sum_U\theta_k$, entonces $t_\theta=Pt_A+(1-P)t_w\;\;t_\theta=Pt_A+(1-P)t_w$

y $t_A=\frac{t_\theta-(1-P)t_W}{P}t_A=\frac{t_\theta-(1-P)t_W}{P}$.

El estimador-$\pi\;\pi$ para $t_\theta\,t_\theta$ es

$$\hat{t}_{\theta,\pi}=\sum_S\frac{Z_k}{\pi_k}$$

El estimador-$\pi\;\pi$ para $t_A\,t_A$ es

$$\hat{t}_{A2,\pi}=\frac{1}{P}\hat{t}_{\theta,\pi}-\frac{(1-P)}{P}t_W$$

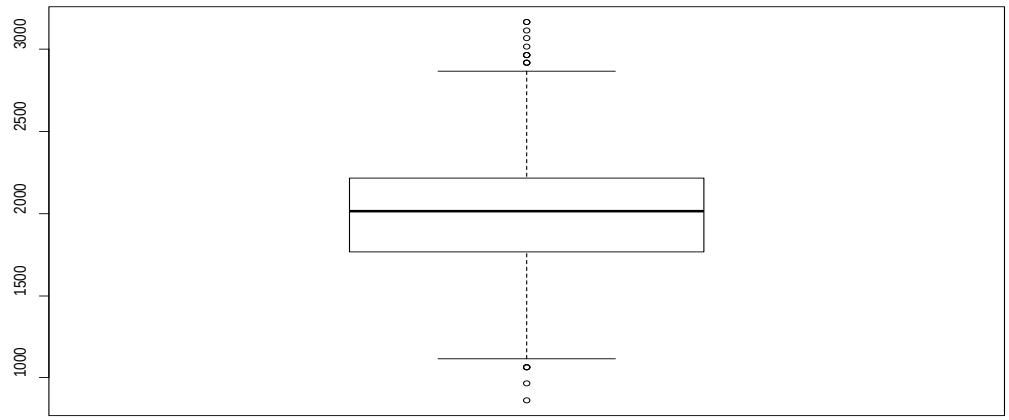
Y su varianza viene dada por

$$V(\hat{t}_{A2,\pi})=\frac{1}{P^2}\left\{\sum\sum_U\Delta_{kl}\breve{\theta}_k\breve{\theta}_l+\sum_U\frac{\theta_k(1-\theta_k)}{\pi_k}\right\}$$

3.3 Simulación para el modelo Mu de respuestas aleatorizadas

```
N<-7000
A<-2000
Y1<-c(rep(0,N-A),rep(1,A))
Y2<-sample(Y1,N,replace=F)
Y<-Y2
B<-3000
W1<-c(rep(0,N-B),rep(1,B))
W2<-sample(W1,N,replace=F)
W<-W2
n<-200
f<-n/N
M<-2000
S<-matrix(rep(0,n*M),nrow=n)
Z<-matrix(rep(2,n*M),nrow=n)
U<-matrix(rep(0,n*M),nrow=n)
TOT.A2<-c(rep(0,M))
TOTTETA.A2<-c(rep(0,M))
P<-0.7
for (j in 1:M){
        S(,j)<-sample(1:N,n,replace=F)
        U(,j)<-runif(n)
        for (k in 1:n){
                if (U(k,j)<P) (Z(k,j)<-Y(S(k,j))) else
               (Z(k,j)<-W(S(k,j)))}
        Z(,j)
        TOTTETA.A2(j)<-sum(Z(,j))/f #Muestreo Aleatorio Simple
        }
TOT.A2<-(1/P)*TOTTETA.A2-((1-P)/P)*B
boxplot (TOT.A2)
summary (TOT.A2)
sqrt(var(TOT.A2))
```

3.4 Resultados de la simulación para el modelo Mu

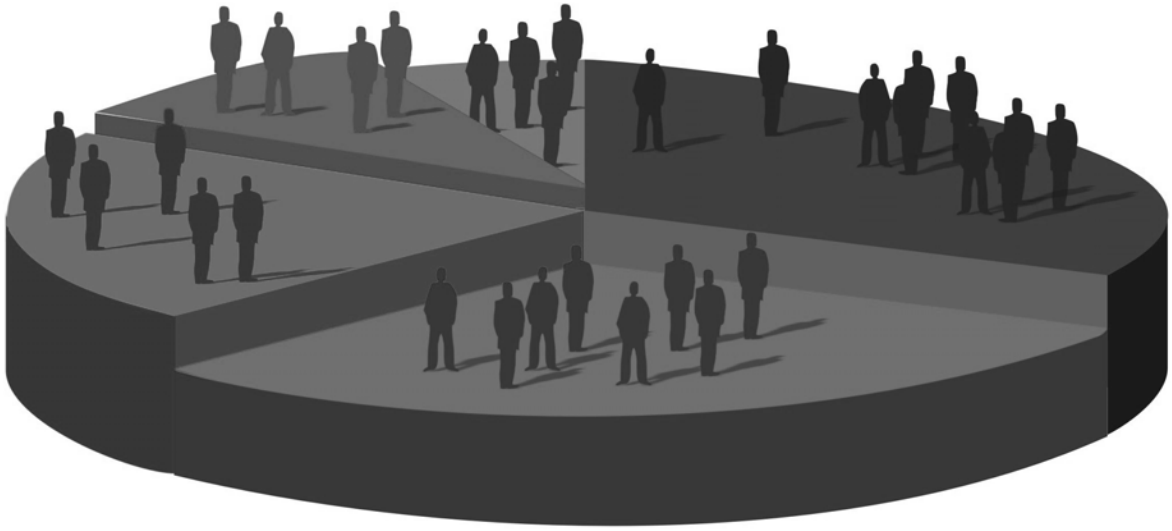



```
Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
                   864.3  1764.0    2014.0    2004.0    2214.0    3164.0

                                    sqrt(var(TOT.A2))=342.7258
```

Conclusión

sqrt(var(tAHPIhat))/ sqrt(var(TOT.A2))/ = 244.46/325.0066=0.7521=75.21%

De manera que el estimador de Conteo de Items de Hussain *et al*. es, en términos de sus desviaciones estándar, 75.21% más eficiente que el estimador MU de Respuestas Aleatorizadas.



Bibliografía

Chaudhuri, A. & Mukherjee, R. (1988), *Randomized Response: Theory and Methods*, Marcel-Decker, New York.
Dalton, D.R & Metzger, M. (1992). 'Integrity testing for personal selection: An unsparing perspective', *Journal of Business Ethics* 12, 147-156.

Droitcour, J. A., 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, in P.P Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz & Sudman, eds, 'Measurement Errors in Surveys', Wiley, New York.

Geurts, M. D. (1980), 'Using a randomized response design to eliminate nonresponse and response bias in business research', *Journal of the Academy of Marketing Science 8*, 83-91.

Hussain, Z., Shah, E. A., Shabbir, J. (2012), 'An alternative Item Count Technique in Sensitive Surveys', Revista Colombiana de Estadística 35, 39-54.

Imai, K., (2010), 'Statistical Inference for Item Count Technique', Tech. Rep., Department of Political Sciences, Princeton University.

Sarndal, C., Swensson, B., Wretman, J., (1992). Model Assisted Survey Sampling. New York, Springer- Verlag.
Soberanis-Cruz, V. H., Ramírez-Valverde, G., Pérez-Elizalde, S., González-Cossio, F. V. (2008), 'Muestreo de Respuestas Aleatorizadas en Poblaciones Finitas: Un Enfoque Unificador', Agrociencia 42, 537-549.

Tracy, D. & Mangat, N. (1996), 'Some development in randomized response sampling during the last decade-a follow up of review by Chaudhuri and Mukerjee', *Journal of Applied Statistical Science* 4, 533-544.
Warner, S. L. (1965), 'Randomized response: A survey technique for eliminating evasive answer bias', *Journal of the American Statistical Association* 60, 63