

# Muestreo con Probabilidades Proporcionales

Curso de Muestreo Probabilístico en Encuestas de Hogares

Andrés Gutiérrez, Ph.D.

CEPAL - Unidad de Estadísticas Sociales

- ① Medidas de tamaño
- ② Diseño de muestreo Poisson (tamaño de muestra aleatorio)
- ③ Diseño de muestreo PPT (con reemplazo)
- ④ Diseño de muestreo  $\pi$ PT (sin reemplazo)

## Motivación

*La estrategia que utiliza un diseño de muestreo aleatorio simple con el estimador de Horvitz-Thompson es óptima bajo ciertas formulaciones, si se tiene un conocimiento a priori de que el comportamiento de la población es simétrico con respecto a los rótulos. En tales casos, la incorporación de información auxiliar no mejora la anterior estrategia.*

Claes-Magnus Cassel (1976)

## Bibliografía y referencias

- Kish, L. (1965) *Survey Sampling*. John Wiley and Sons.
- Cochran, W. G. (1977) *Sampling Techniques*. John Wiley and Sons.
- Särndal, et. al. (2003) *Model-assisted Survey Sampling*. Springer.
- Gutiérrez, H. A. (2016) *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Ediciones de la U.
- Gutiérrez, H. A. (2017) *TeachingSampling. R package*.

## Medidas de tamaño

## Probabilidades desiguales

- En cuestión de precisión, se puede tener una mayor ganancia cuando se utilizan diseños de muestreo con probabilidades desiguales.
- En la mayoría de los casos prácticos, la característica de interés no presenta un comportamiento uniforme con respecto a los rótulos de la población.

## Probabilidades desiguales

Cuando el marco de muestreo disponible para la selección de la muestra contiene una característica auxiliar continua disponible para todos los elementos de la población  $x_k \quad \forall k \in U$ , es posible utilizar diseños de muestreo que implementen métodos de selección cuyas probabilidades de selección o inclusión, dependiendo del caso, sean proporcionales a  $x_k$ .

Diseño de muestreo Poisson (tamaño de muestra aleatorio)

## Diseño de muestreo de Poisson

Siendo  $\pi_k$  un número positivo, tal que  $0 < \pi_k \leq 1$ , que representa la probabilidad de inclusión del  $k$ -ésimo elemento, el diseño de muestreo Poisson se define de la siguiente manera

$$p(s) = \prod_{k \in s} \pi_k \prod_{k \notin s} (1 - \pi_k) \quad \text{para todo } s \in Q \quad (1)$$

con  $Q$ , el soporte que contiene a todas las posibles muestras sin reemplazo.

## Posibles muestras

- En nuestra población ejemplo

$$U = \{\text{Yves, Ken, Erik, Sharon, Leslie}\}$$

Las probabilidades de inclusión  $\pi_k$  son 0.2, 0.5, 0.7, 0.5 y 0.9, respectivamente.

- Las posibles muestra pueden ser de tamaño 0, 1, 2, 3, 4 ó 5.

## Posibles muestras (n=0)

La probabilidad de la muestra de tamaño 0 es

$$(1 - 0.2) \times (1 - 0.5) \times (1 - 0.7) \times (1 - 0.5) \times (1 - 0.9) = 0.006$$

## Posibles muestras (n=1)

s	p(s)
Yves	0.0015
Ken	0.006
Erik	0.014
Sharon	0.006
Leslie	0.054
Total	0.0815

## Posibles muestras (n=2)

s	p(s)
Yves, Ken	0.0015
Yves, Erik	0.0035
Yves, Sharon	0.0015
Yves, Leslie	0.0135
Ken, Erik	0.014
Ken, Sharon	0.006
Ken, Leslie	0.054
Erik, Sharon	0.014
Erik, Leslie	0.126
Sharon, Leslie	0.054
Total	0.288

## Posibles muestras (n=3)

s	p(s)
Yves, Ken, Erik	0.0035
Yves, Ken, Sharon	0.0015
Yves, Ken, Leslie	0.0135
Yves, Erik, Sharon	0.0035
Yves, Erik, Leslie	0.0315
Yves, Sharon, Leslie	0.0135
Ken, Erik, Sharon	0.014
Ken, Erik, Leslie	0.126
Ken, Sharon, Leslie	0.054
Erik, Sharon, Leslie	0.126
Total	0.387

## Posibles muestras (n=4)

s	p(s)
Yves, Ken, Erik, Sharon	0.0035
Yves, Erik, Sharon, Leslie	0.0315
Yves, Ken, Erik, Leslie	0.0315
Yves, Ken, Sharon, Leslie	0.0135
Ken, Erik, Sharon, Leslie	0.126
Total	0.206

## Posibles muestras (n=5)

- Finalmente, la muestra de tamaño 5,  $\{Yves, Ken, Erik, Sharon, Leslie\}$ , tiene probabilidad 0.0315.
- Nótese que la suma de todas las posibles muestras es  $\sum p(s) = 1$ .

## Inclusión forzosa

- Existen elementos de la población que deben ser observados obligatoriamente en la muestra.
- En estos casos el valor de la probabilidad de inclusión de estos elementos es igual a uno ( $\pi_k = 1$ ).
- Al subgrupo poblacional cuyos elementos tienen probabilidad de inclusión igual a uno, se le conoce como subgrupo de **inclusión forzosa**.

## Algoritmo de selección

La selección de una muestra con diseño de muestreo Poisson se realiza mediante un algoritmo secuencial definido así:

- Fijar para cada  $k \in U$  el valor de la probabilidad de inclusión  $\pi_k$  tal que  $0 < \pi_k \leq 1$ .
- Obtener  $\varepsilon_k$  para  $k \in U$  como  $N$  realizaciones independientes de una variable aleatoria con distribución uniforme en el intervalo  $[0, 1]$ .
- El elemento  $k$ -ésimo pertenece a la muestra con probabilidad  $\pi_k$ . Es decir, si  $\varepsilon_k < \pi$  el individuo  $k$ -ésimo es seleccionado.

## Algoritmo de selección

- Suponga que el individuo **Erik** debe estar en la muestra seleccionada, es decir,  $\pi_{Erik} = 1$ .
- Además, el vector de probabilidades de inclusión para cada elemento de la población está dado por  $(0.5, 0.2, 1, 0.9, 0.5)$ .
- Existen  $\binom{1}{1}2^4 = 16$  posibles muestras.

## Algoritmo de selección

k	Nombre	pik
1	Yves	0.5
2	Ken	0.2
3	Erik	1.0
4	Sharon	0.9
5	Leslie	0.5

## Algoritmo de selección

k	Nombre	pik	ek	Ik
1	Yves	0.5	0.550	0
2	Ken	0.2	0.041	1
3	Erik	1.0	0.750	1
4	Sharon	0.9	0.649	1
5	Leslie	0.5	0.494	1

## El estimador del total

El estimador de Horvitz-Thompson y su varianza estimada están dados por:

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} \quad (2)$$

$$\widehat{Var}_{PO}(\hat{t}_{y,\pi}) = \sum_S (1 - \pi_k) \left( \frac{y_k}{\pi_k} \right)^2 \quad (3)$$

respectivamente.

## ¿Cómo definir los $\pi_k$ ?

*Suponiendo un tamaño de muestra fijo, bajo un diseño de muestreo Poisson, la varianza del estimador de Horvitz-Thompson se minimiza cuando*

$$\pi_k = n \frac{y_k}{\sum_U y_k} \quad (4)$$

## ¿Cómo definir los $\pi_k$ ?

- El anterior resultado es una ambigüedad puesto que con esa escogencia de las probabilidades de inclusión se asume que la característica de interés es conocida para toda la población.
- Si lo anterior sucede, no existiría la necesidad de estimar  $t_y$

## ¿Cómo definir los $\pi_k$ ?

- Si el marco de muestreo tiene la **virtud** de adjuntar información auxiliar continua, por medio de una característica de interés  $x_k$  se puede hacer mínima la varianza.
- En otras palabras, es necesario conocer el vector de características auxiliares  $x_1, x_2, \dots, x_N$ , antes de realizar el muestreo.
- $x_k$  debe estar muy bien correlacionada con la variable de interés.
- Las probabilidades de inclusión se definen así:

$$\pi_k = n \frac{x_k}{\sum_U x_k} \quad (5)$$

## Práctica en R

```
library(TeachingSampling)
library(dplyr)
data("BigCity")

Hogares <- BigCity %>% group_by(HHID) %>%
  summarise(Ingreso = sum(Income),
            Gasto = sum(Expenditure),
            EdadMedia = mean(Age),
            Personas = n())
```

## Práctica en R

```
head(Hogares)
```

HHID	Ingreso	Gasto	EdadMedia	Personas
idHH00001	2775	2442	27	5
idHH00002	1492	1084	19	5
idHH00003	4280	2441	38	4
idHH00004	2200	1851	30	4
idHH00005	3119	3068	32	5
idHH00006	675	1098	25	5

## Práctica en R

```
attach(Hogares)
N <- dim(Hogares) [1]
n <- 2000
pik <- n * Personas / sum(Personas)
which(pik > 1)

## integer(0)
sum(pik)

## [1] 2000
```

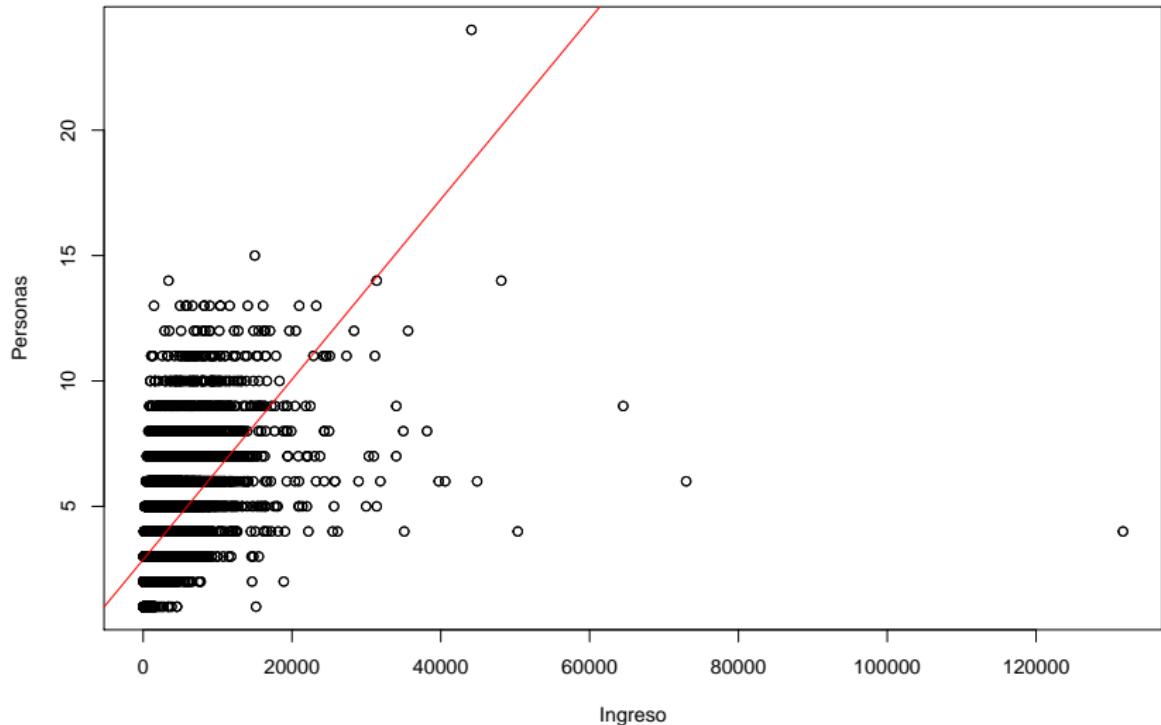
## Práctica en R

```
matriz <- cbind(pik, Ingreso, Gasto, Personas)
cor(matriz)
```

	pik	Ingreso	Gasto	Personas
pik	1.00	0.56	0.64	1.00
Ingreso	0.56	1.00	0.74	0.56
Gasto	0.64	0.74	1.00	0.64
Personas	1.00	0.56	0.64	1.00

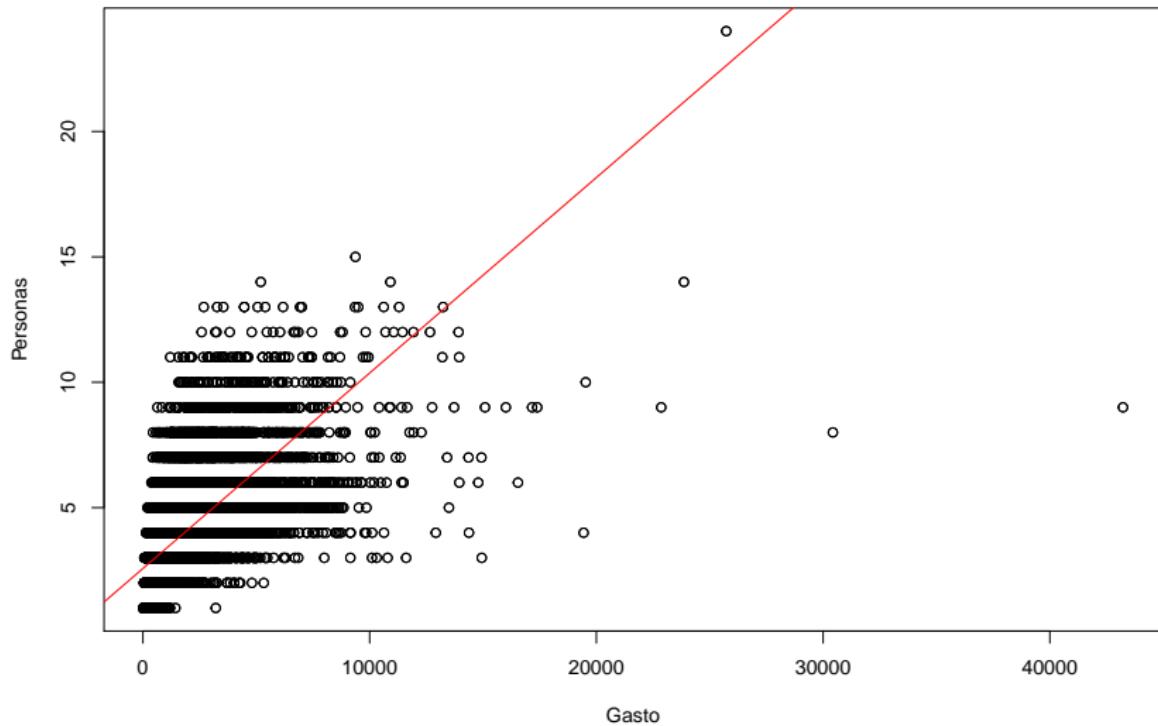
## Práctica en R

```
plot(Personas ~ Ingreso)  
abline(lm(Personas ~ Ingreso), col=2)
```



## Práctica en R

```
plot(Personas ~ Gasto)
abline(lm(Personas ~ Gasto), col=2)
```



## Práctica en R

```
sam <- S.PO(N, pik)
muestra <- Hogares[sam,]
n.s <- dim(muestra)[1]
n.s

## [1] 1972
```

## Práctica en R

```
attach(muestra)  
head(muestra)
```

HHID	Ingreso	Gasto	EdadMedia	Personas
idHH00031	2049	1732	35	5
idHH00038	1008	297	38	3
idHH00044	1577	2195	27	6
idHH00045	939	302	52	2
idHH00094	4121	2000	18	3
idHH00148	7797	3385	22	9

## Práctica en R

```
pik.s <- pik[sam]  
estima <- data.frame(Ingresa, Gasto, Personas)  
E.PO(estima, pik.s)
```

	N	Ingresa	Gasto	Personas
Estimation	41025.9	89041391.80	55786147.31	148162.3
Standard Error	1062.8	3545096.70	1519582.83	3233.0
CVE	2.6	3.98	2.72	2.2
DEFF	Inf	0.62	0.61	2.8

Diseño de muestreo PPT (con reemplazo)

## Razonamiento

- En un diseño de muestreo con reemplazo, los valores óptimos de las probabilidades de selección para cada elemento de la población tendrían que estar dados por

$$p_k = \frac{y_k}{t_y}.$$

- Con esta escogencia, el estimador de Hansen-Hurwitz estimaría al total poblacional de la característica de interés con varianza nula.

## Razonamiento

Inclusive, el tamaño de muestra necesario para obtener una estimación con sesgo y varianza nula sería de  $m = 1$ .

$$\begin{aligned}\hat{t}_{y,p} &= \frac{1}{m} \sum_{i=1}^m \frac{y_{k_i}}{p_{k_i}} \\ &= \frac{1}{1} \sum_{i=1}^1 \frac{y_{k_i}}{p_{k_i}} \\ &= \frac{y_{k_i}}{p_{k_i}} \\ &= t_y \frac{y_{k_i}}{y_{k_i}} = t_y\end{aligned}$$

## Razonamiento

- Desde el punto de vista práctico sería ambigüedad la escogencia de las anteriores probabilidades de selección.
- Sin embargo, si el marco de muestreo contiene el valor de  $x_k$  bien relacionada con la característica de interés  $y_k$ , es posible estimar el parámetro de interés con una varianza pequeña.
- Entre mejor correlación exista entre  $y_k$  y  $x_k$  menor varianza tendrá el estimador de Hansen-Hurwitz.

## Selección: método acumulativo total

Este algoritmo consiste en  $m$  selecciones independientes de tamaño 1, tal que:

- Sea

$$p_k = \frac{x_k}{t_x} \quad (6)$$

- Sea

$$t_k = \sum_{l=1}^k x_l \quad (7)$$

con  $t_0 = 0$

- Obtener  $\varepsilon$  como una realización de una variable aleatoria con distribución uniforme en el intervalo  $(0,1)$ .
- Seleccionar el  $k$ -ésimo elemento si  $t_{k-1} < \varepsilon * t_x \leq t_k$ .

## Algoritmo de selección

```
U <- c("Yves", "Ken", "Erik", "Sharon", "Leslie")
x <- c(52, 60, 75, 100, 50)
sum(x)
```

```
## [1] 337
```

```
pk <- x / sum(x)
pk
```

```
## [1] 0.15 0.18 0.22 0.30 0.15
```

```
sum(pk)
```

```
## [1] 1
```

## Algoritmo de selección ( $i=1$ )

Seleccionemos una muestra de  $m = 3$  elementos. Para la primera selección tenemos  $\varepsilon_1 = 0.58$  y  $\varepsilon * t_x = 195$ .

k	Nombre	xk	tk-1	tk
1	Yves	52	0	52
2	Ken	60	52	112
3	Erik	75	112	187
4	Sharon	100	187	287
5	Leslie	50	287	337

El primer elemento seleccionado es **Sharon**.

## Algoritmo de selección ( $i=1$ )

Para la segunda selección tenemos  $\varepsilon_2 = 0.13$  y  $\varepsilon * t_x = 44$ .

k	Nombre	xk	tk-1	tk
1	Yves	52	0	52
2	Ken	60	52	112
3	Erik	75	112	187
4	Sharon	100	187	287
5	Leslie	50	287	337

El segundo elemento seleccionado es **Yves**.

## Algoritmo de selección ( $i=3$ )

Para la tercera selección tenemos  $\varepsilon_3 = 0.65$  y  $\varepsilon * t_x = 219$ .

k	Nombre	xk	tk-1	tk
1	Yves	52	0	52
2	Ken	60	52	112
3	Erik	75	112	187
4	Sharon	100	187	287
5	Leslie	50	287	337

El tercer elemento seleccionado es **Sharon**.

## Estimador del total poblacional

Sea  $x_k$  el valor de una característica auxiliar continua, el estimador de Hansen-Hurwitz del total poblacional  $t_y$  y su varianza estimada están dados por:

$$\hat{t}_{y,p} = \frac{t_x}{m} \sum_{i=1}^m \frac{y_{ki}}{x_{ki}} \quad (8)$$

$$\widehat{Var}_{PPT}(\hat{t}_{y,p}) = \frac{1}{m(m-1)} \sum_{i=1}^m \left( \frac{y_i}{p_i} - \hat{t}_{y,p} \right)^2 \quad (9)$$

respectivamente.

## Estimador del total poblacional

*Para el diseño de muestreo PPT, el estimador de Hansen-Hurwitz del total de la característica de información auxiliar reproduce ese total con varianza nula*

Nótese que

$$\hat{t}_{x,p} = \frac{1}{m} \sum_{k \in S} \frac{x_k}{p_k} = \frac{1}{m} \sum_{k \in S} \frac{x_k}{x_k/t_x} = \frac{1}{m} \sum_{k \in S} t_x = t_x$$

## Eficiencia: la regla de oro

*Para que la inferencia basada en el diseño de muestreo arroje estimaciones que sean de varianza mínima e insesgadas, las probabilidades de inclusión (o selección) que arroje el diseño de muestreo deben ser directamente proporcionales a la característica de interés.*

$$\pi_k \propto y_k$$

$$p_k \propto y_k$$

## Eficiencia

La resta de la varianza de la estrategia aleatoria simple con la varianza de la estrategia PPT da como resultado la siguiente expresión:

$$Var_{MRAS}(\hat{t}_{y,p}) - Var_{PPT}(\hat{t}_{y,p}) = \frac{N^2}{m} Cov \left( x, \frac{y^2}{x} \right) \quad (10)$$

## Eficiencia

Para que la estrategia de muestreo PPT sea más eficiente, es necesario que:

- ① La correlación entre  $\left(x, \frac{y^2}{x}\right)$  sea positiva.
- ② La razón  $\frac{y_k}{x_k}$  permanezca constante para todo  $k \in U$ .
- ③ El valor de  $\beta_0$  debe ser pequeño si el siguiente modelo de regresión se asume

$$y_k = \beta_0 + \beta_1 x_k + E_k \quad (11)$$

# Práctica en R

```
attach(Hogares)
N <- nrow(Hogares)
m <- 2000

(N^2 / m) * cov(Personas, (Ingreso^2 / Personas))

## [1] 3443098156019
(N^2 / m) * cov(Personas, (Gasto^2 / Personas))

## [1] 1100113721196
```

## Práctica en R

```
cor(Personas, (Ingreso2 / Personas))
```

```
## [1] 0.068
```

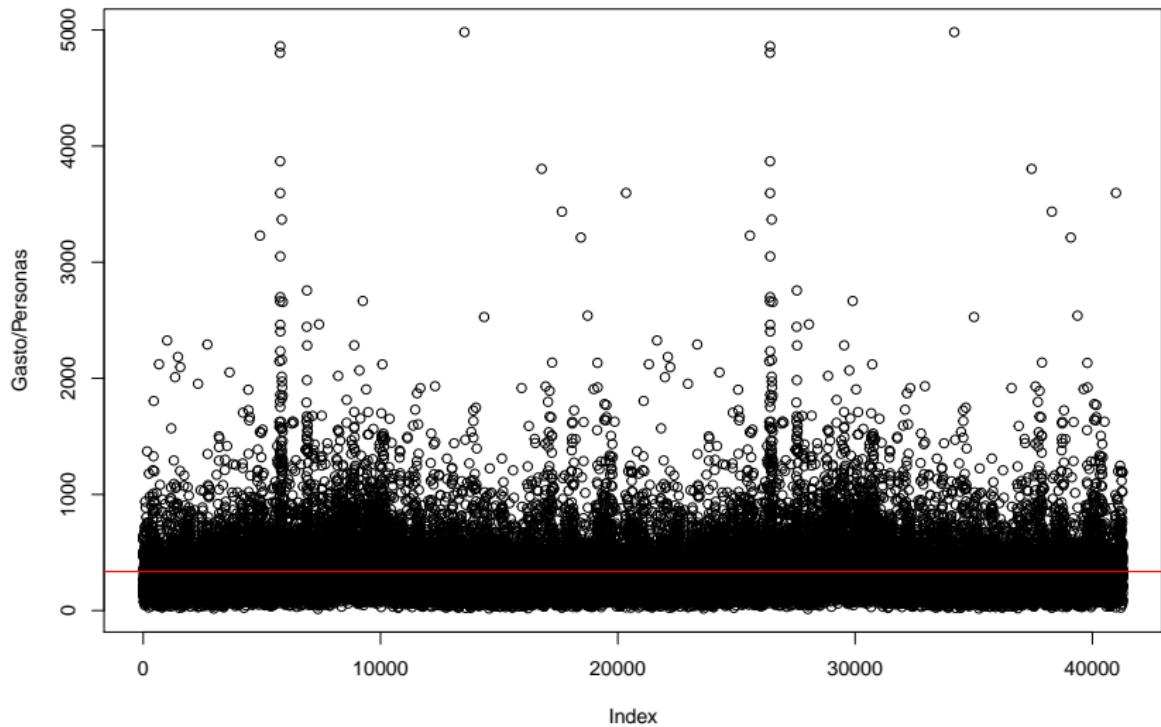
```
cor(Personas, (Gasto2/Personas))
```

```
## [1] 0.26
```

- Para la estimación del total de gastos e ingresos declarados, sí se tiene una mayor eficiencia.

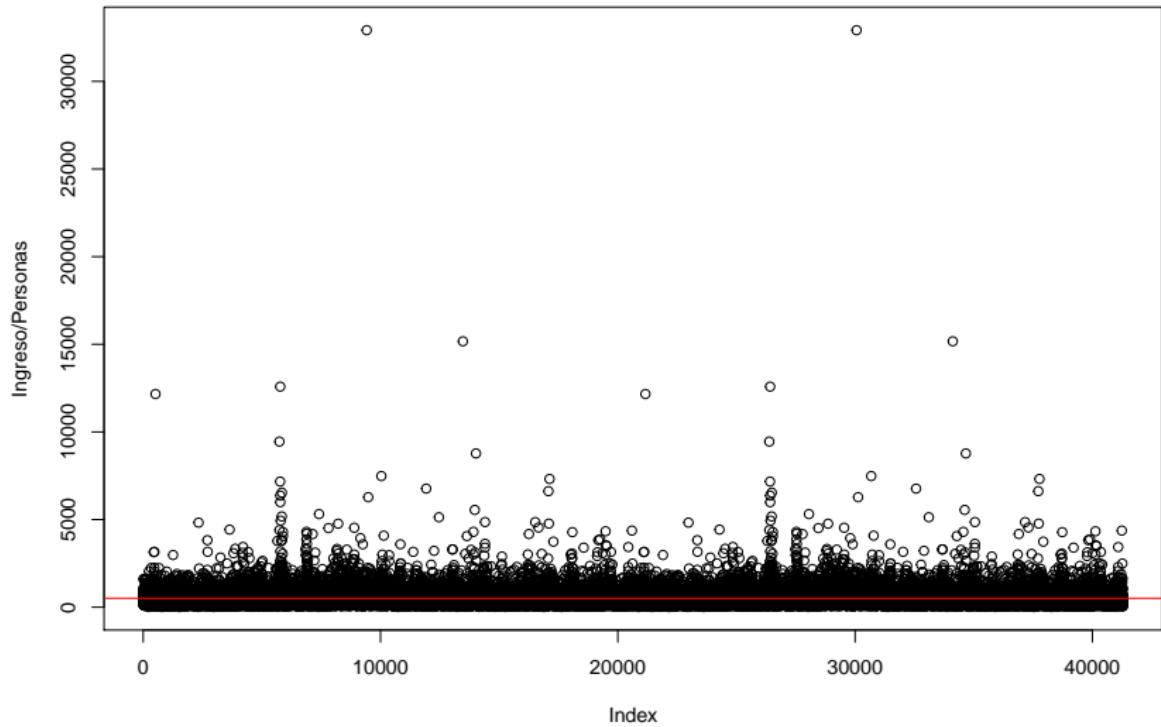
## Práctica en R

```
plot(Gasto/Personas)
abline(h = mean(Gasto/Personas), col = 2)
```



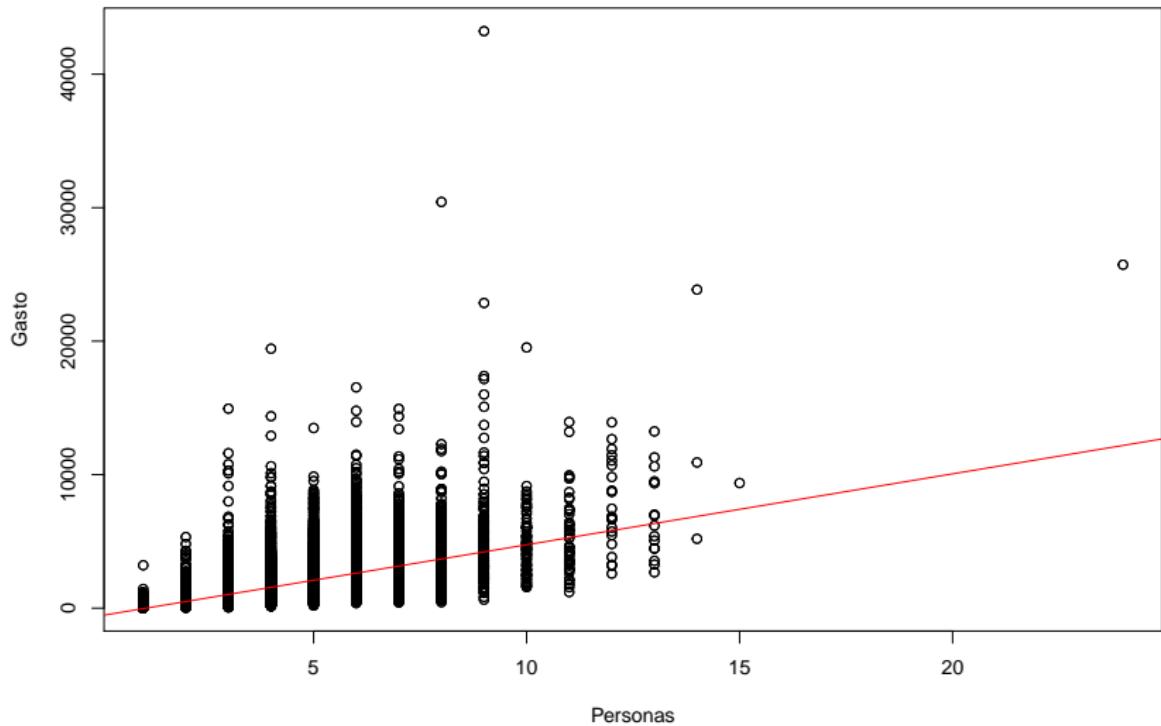
## Práctica en R

```
plot(Ingreso/Personas)
abline(h = mean(Ingreso/Personas), col = 2)
```



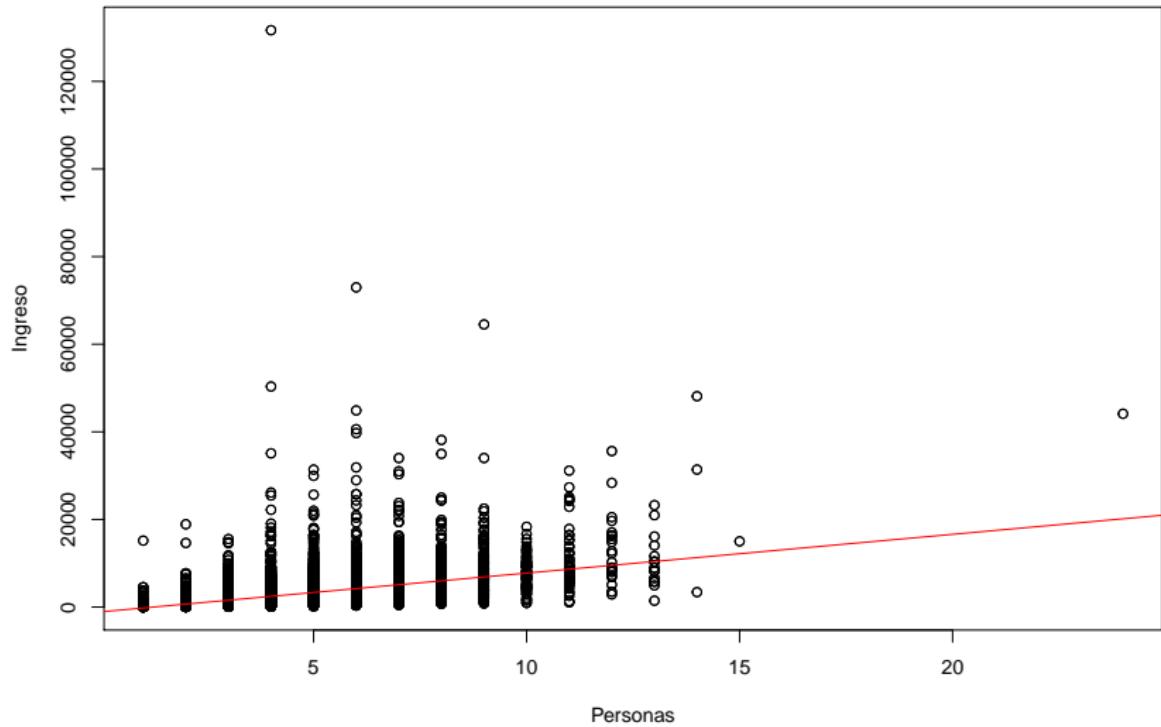
# Práctica en R

```
plot(Gasto ~ Personas)
abline(lm(Gasto ~ Personas), col=2)
```



## Práctica en R

```
plot(Ingresa ~ Personas)  
abline(lm(Ingresa ~ Personas), col=2)
```



# Práctica en R

```
M.I <- lm(Gasto ~ Personas)
summary(M.I)

##
## Call:
## lm(formula = Gasto ~ Personas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4076   -557   -135    245  39010 
##
## Coefficients:
##             Estimate Std. Error t value     Pr(>|t|)    
## (Intercept) -566.95     12.70  -44.6 <0.0000000000000002 ***
## Personas      531.47      3.11   170.9 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1170 on 41288 degrees of freedom
## Multiple R-squared:  0.414, Adjusted R-squared:  0.414 
## F-statistic: 2.92e+04 on 1 and 41288 DF,  p-value: <0.0000000000000002
```

# Práctica en R

```
M.E <- lm(Ingreso ~ Personas)
summary(M.E)

##
## Call:
## lm(formula = Ingreso ~ Personas)
##
## Residuals:
##     Min      1Q Median      3Q     Max 
## -8955  -1007   -239    451 129232 
##
## Coefficients:
##             Estimate Std. Error t value     Pr(>|t|)    
## (Intercept) -1091.65     26.05  -41.9 <0.0000000000000002 ***
## Personas      884.88      6.38   138.7 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2400 on 41288 degrees of freedom
## Multiple R-squared:  0.318, Adjusted R-squared:  0.318 
## F-statistic: 1.92e+04 on 1 and 41288 DF,  p-value: <0.0000000000000002
```

## Práctica en R

```
pk <- Personas / sum(Personas)
sam <- S.PPS(m, Personas)
muestra <- Hogares[sam,]
attach(muestra)
```

## Práctica en R

```
head(muestra)
```

HHID	Ingreso	Gasto	EdadMedia	Personas
idHH14325	872	208	41	2
idHH15577	150	43	80	1
idHH23877	720	615	18	3
idHH32724	1920	901	35	3
idHH08205	4057	6222	25	4
idHH17972	3652	2634	26	5

## Práctica en R

```
pk.s <- pk[sam]  
estima <- data.frame(Ingresa, Gasto, Personas)  
E.PPS(estima, pk.s)
```

	N	Ingresa	Gasto	Personas
Estimation	41391.9	87807366.09	57484147.85	150266
Standard Error	555.1	1736515.90	985131.10	0
CVE	1.3	1.98	1.71	0
DEFF	Inf	0.33	0.29	0

Diseño de muestreo  $\pi$ PT (sin reemplazo)

## Características

- Utilizar un esquema de muestreo con probabilidades proporcionales a alguna característica de información auxiliar puede resultar en ganancia de precisión.
- Utilizar una estrategia de muestreo que contemple reemplazo es menos eficiente

## Características

Lo deseable es poder implementar una estrategia de muestreo que:

- ① Contemple un diseño de muestreo sin reemplazo.
- ② Sea de probabilidades proporcionales.
- ③ Sea de tamaño muestral fijo.

## Probabilidades de inclusión

El diseño de muestreo  $\pi$ PT se basa en la construcción de probabilidades de inclusión que obedezcan la siguiente relación:

$$\pi_k = \frac{nx_k}{t_x} \quad 0 < \pi_k \leq 1 \quad (12)$$

## Probabilidades de inclusión

Un primer paso para el cálculo de las probabilidades de inclusión es aplicar la expresiones anteriores.

```
n <- 4
x <- c(52, 60, 75, 100, 50)
pik <- n * x / sum(x)
pik
## [1] 0.62 0.71 0.89 1.19 0.59
```

## Probabilidades de inclusión

- El cuarto elemento de la población, correspondiente a **Sharon** es un elemento de inclusión forzosa
- **Sharon** debe estar presente en todas las posibles muestras.
- El siguiente paso es separar a **Sharon** de los restantes elementos y proseguir con el cálculo de las probabilidades de inclusión.

## Probabilidades de inclusión

```
n <- 3
x <- c(52, 60, 75, 50)
pik <- n * x / sum(x)
pik
## [1] 0.66 0.76 0.95 0.63
```

## Probabilidades de inclusión

Por tanto el vector de probabilidades de inclusión para toda la población  $U$  está dado por

$$\pi = (\underbrace{0.6582278}_{\text{Yves}}, \underbrace{0.7594937}_{\text{Ken}}, \underbrace{0.9493671}_{\text{Erik}}, \underbrace{1.0000}_{\text{Sharon}}, \underbrace{0.6329114}_{\text{Leslie}})'$$

## Algoritmo de selección

Con lo anterior se busca que:

- ① El algoritmo de selección de muestras bajo este diseño sea de fácil implementación computacional.
- ② Las probabilidades de inclusión de segundo orden sean positivas,  $\pi_{kl} > 0$ . De lo contrario el estimador de la varianza podría ser sesgado.
- ③ El cálculo de estas probabilidades de inclusión de segundo orden,  $\pi_{kl}$ , sea sencillo.
- ④  $\Delta_{kl} < 0 \quad \forall k \neq l$  para que la estimación de la varianza no sea negativa.

## Selección: método de Sunter

- ① Ordenar descendente la población de acuerdo con los valores que toma la característica de información auxiliar  $x_k$ .
- ② Realizar  $\xi_k \sim U(0, 1)$ .
- ③ Para  $k = 1$ , el primer elemento de la lista ordenada es incluido en la muestra si y solamente si  $\xi_1 < \pi_1$ .
- ④ Para  $k \geq 2$ , el  $k$ -ésimo elemento de la lista ordenada es incluido en la muestra si y solamente si

$$\xi_k \leq \frac{n - n_{k-1}}{n - \sum_{i=1}^{k-1} \pi_i} \pi_k$$

donde  $n_{k-1}$  representa el número de elementos que ya han sido seleccionados al final del paso  $k - 1$ .

## Selección: método de Sunter

k	Nombre	xk	Pik	ek	wk	Ik	nk
4	Sharon	100	0.89	0.27	-	1	0
3	Erik	75	0.67	0.49	0.63	1	1
2	Ken	60	0.53	0.74	0.37	0	2
1	Yves	52	0.46	0.43	0.51	1	2
5	Leslie	50	0.45	0.65	0.00	0	3

## Estimador del total poblacional

El estimador de Horvitz-Thompson y su varianza estimada están dados por:

$$\hat{t}_{y,\pi} = \sum_S \frac{y_k}{\pi_k} \quad (13)$$

$$\widehat{Var}_{\pi PT}(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum_S \sum_{kl} \frac{\Delta_{kl}}{\pi_{kl}} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \quad (14)$$

respectivamente.

## Propiedad de calibración

*Para el diseño de muestreo  $\pi PT$ , el estimador de Horvitz-Thompson del total de la característica de información auxiliar reproduce ese total con varianza nula*

$$\hat{t}_{x,\pi} = \sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in S} t_x \frac{1}{n} = t_x$$

## Aproximación de la varianza

Tillé (2002) menciona que:

- *Tiene la convicción de que las probabilidades de inclusión de segundo orden no son usadas para nada.*
- *En la práctica el uso de las probabilidades de inclusión de segundo orden es muchas veces irreal porque son muy difíciles de calcular computacionalmente.*
- *Para hacerlo se debe computar  $n^2$  términos que deben ser sumados para calcular la estimación.*

## Aproximación de la varianza

La aproximación de la varianza del estimador de Horvitz-Thompson está dada por

$$\widehat{Var}(\hat{t}_{y,\pi}) = \sum_{k \in S} \frac{c_k}{\pi_k^2} (y_k - \hat{y}_k^*)^2 \quad (15)$$

donde

$$\hat{y}_k^* = \pi_k \frac{\sum_{l \in S} c_l y_l / \pi_l}{\sum_{l \in S} c_l} \quad (16)$$

y

$$c_k = (1 - \pi_k) \frac{n}{(n - 1)} \quad (17)$$

## Práctica en R

```
attach(Hogares)
N <- dim(Hogares)[1]
n <- 2000

res <- S.piPS(n, Personas)
sam <- res[,1]
muestra <- Hogares[sam,]
attach(muestra)
```

## Práctica en R

```
head(muestra)
```

HHID	Ingreso	Gasto	EdadMedia	Personas
idHH40586	6630	3269	13	13
idHH35809	14061	11296	18	13
idHH19941	6630	3269	13	13
idHH15164	14061	11296	18	13
idHH13692	10343	4468	24	13
idHH12224	10404	6182	20	13

## Práctica en R

```
pik.s <- res[, 2]  
estima <- data.frame(Ingresa, Gasto, Personas)  
E.piPS(estima, pik.s)
```

	N	Ingresa	Gasto	Personas
Estimation	42522.1	87899255.25	55541041.68	150266
Standard Error	557.6	1945423.34	902179.92	0
CVE	1.3	2.21	1.62	0
DEFF	Inf	0.42	0.27	0

¡Gracias!

Andrés Gutiérrez

*Experto Regional en Estadísticas Sociales*

*Division de Estadísticas*

*Email: andres.GUTIERREZ@cepal.org*