

# **NUEVAS TECNICAS PARA LA ESTIMACION DE PARAMETROS EN ENCUESTAS POR MUESTREO**

## **APLICACIÓN EN S.A.S.**

lunes, 29 de agosto de 2011

# CONTENIDO

- Introducción
- Esquema de simulación
- Diseño MAS
- Diseño ESTMAS
- Diseño MAS<sup>2</sup>
- Conclusiones

# INTRODUCCION

- ❑ Alternativas para la estimación de totales, razones y coeficientes de variación:
  - **Programación manual de las formulas.**
  - ✓ Errores de muestreo observados, no estimados.
  - ✓ Dificultad en la programación que puede inducir a errores.
  - ✓ “Programación individual”.
  - **Procedimientos en SAS**
  - ✓ Facilidad en la programación y generación masiva de resultados.
  - ✓ “Programación grupal”
  - ✓ Errores de muestreo estimados (para diseños en etapas).

# OBJETIVO

1. Comparar los resultados obtenidos en la estimación de totales, razones y errores de muestreo para los diseños:
  - ✓ M.A.S
  - ✓ EST-M.A.S
  - ✓ M.A.S<sup>2</sup>
2. Mostrar como se pueden generar estimaciones de los parámetros utilizando los procedimientos de S.A.S.

# SIMULACION

1. Para cada diseño de muestreo se simula un universo para el estudio de totales y razones.
2. Se selecciona una muestra
3. Se generan las estimaciones para totales, razones y cve's por las dos alternativas estudiadas.
4. Se repiten los pasos 2-4 conservando los resultados de la estimación.

# UNIVERSO M.A.S.

- **SIMULACION DEL UNIVERSO DE ESTUDIO**

- `DATA UNIVERSO;`
- `DO I=1 TO 8000;`
- `XK=2*rangam(12345,100); /*CHI CUADRADO DE MEDIA 200*/`
- `ALEA1=RANUNI(23);`
- `ALEA2=RANUNI(45);`
- `YK=0; ZK=0;`
- `IF ALEA1<=0.9 THEN ZK=1;`
- `IF ALEA2<=0.1 THEN YK=1; /*R APROXIMADAMENTE DE 11%*/`
- `IF YK=1 AND ZK=0 THEN YK=0;`
- `PEGA=1;`
- `OUTPUT;`
- `END;`
- `KEEP I XK YK ZK PEGA;`
- `RUN;`

- **SELECCIÓN DE LA MUESTRA UTILIZANDO SURVEYSELECT**

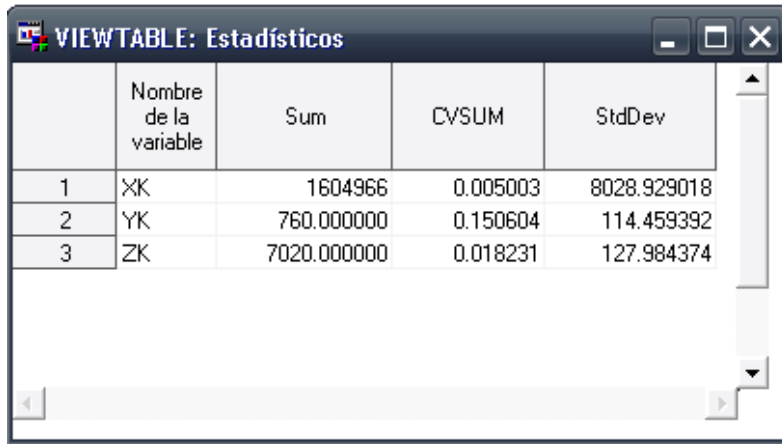
- `PROC SURVEYSELECT DATA=UNIVERSO METHOD=SRS N=400 OUT=MUESTRA_MAS  
OUTSIZE;`
- `RUN;`

# ESTIMACIONES CON S.A.S.

- `PROC SURVEYMEANS DATA=MUESTRA_MAS N=8000 SUM CVSUM;`
- `weight FEXP;`
- `ratio YK/ ZK;`
- `VAR XK YK ZK;`
- `ODS OUTPUT Statistics=TOTALES RATIO=RAZON;`
- `RUN;`
  
- Se debe especificar el tamaño poblacional (en caso de conocerlo) en la sintaxis.
- **SUM** calcula la estimación del total.
- **CVSUM** calcula el cve para el total.
- Tabla TOTALES guarda las estimaciones para el total y su cve.
- Tabla RAZON guarda las estimaciones para la razón y su error estándar.

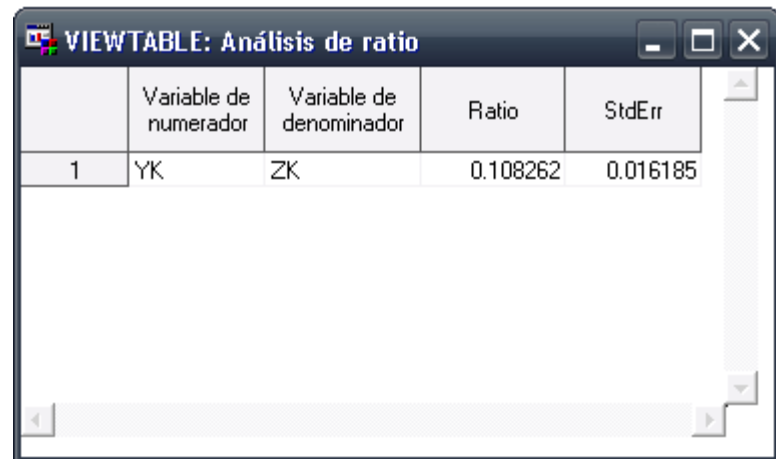
# TABLAS DE RESULTADOS

## Tabla Totales



	Nombre de la variable	Sum	CVSUM	StdDev
1	XK	1604966	0.005003	8028.929018
2	YK	760.000000	0.150604	114.459392
3	ZK	7020.000000	0.018231	127.984374

## Tabla Razon



	Variable de numerador	Variable de denominador	Ratio	StdErr
1	YK	ZK	0.108262	0.016185

*The SURVEYMEANS Procedure*

## Syntax

The following statements are available in PROC SURVEYMEANS.

**PROC SURVEYMEANS** < options > < statistic-keywords > ;

**BY** variables ;

**CLASS** variables ;

**CLUSTER** variables ;

**DOMAIN** variables < variable\*variable  
variable\*variable\*variable ... > ;

**RATIO** < 'label' > variables / variables ;

**STRATA** variables < / option > ;

**VAR** variables ;

**WEIGHT** variable ;

Además

....

# RESULTADOS

- ✓ Para cada muestra se calcula la diferencia entre las estimaciones del total y de la razón, así como el cociente entre los coeficientes de variación estimados.
- ✓ Se obtiene un promedio para cada tamaño de muestra evaluado.

## Resumen comparación de estimaciones

n	diff total	dif razón	cociente cve(total)	cociente cve(razón)
200	-2.00E-10	-1.94E-18	1	1
400	-3.73E-10	-1.39E-18	1	1
600	2.33E-10	1.94E-18	1	1
800	-4.98E-10	-8.33E-19	1	1

*500 muestras para cada tamaño de muestra*

Conclusión: Resultados exactamente iguales.

- Si no se especifica la opción N=8000 en el procedimiento surveymeans los resultados son los siguientes:

Resumen comparación de estimaciones

n	diff total	dif razón	cociente cve(total)	cociente cve(razón)
200	1.72E-10	-2.08E-18	1.0127	1.0127
400	-1.63E-10	-1.11E-18	1.0260	1.0260
600	8.85E-11	-2.78E-18	1.0398	1.0398
800	-3.96E-10	0.00E+00	1.0541	1.0541

*500 muestras para cada tamaño de muestra*

Conclusión: Totales y razones iguales pero errores de muestreo levemente superiores con el procedimiento surveymeans.

# DISEÑO ESTMAS

- **SIMULACION DEL UNIVERSO DE ESTUDIO**
- **4 estratos**
- **12819 individuos**
- **DATA UNIVERSO;**
- **DO** ESTRATO=1 **TO** 4;
- NH=ROUND((3000+1500\*(ranuni(123)-0.5)));
- do ind=1 to NH;
- XK=2\*rangam(12345,100); /\*CHI CUADRADO DE MEDIA 200\*/
- ALEA1=RANUNI(234);
- ALEA2=RANUNI(456);
- YK=0;ZK=0;
- IF ALEA1<=0.9 THEN ZK=1;
- IF ALEA2<=0.1 THEN YK=1;
- IF YK=1 AND ZK=0 THEN YK=0;
- PEGA=1;
- OUTPUT;
- END;
- END;
- KEEP ESTRATO NH XK YK ZK PEGA;
- **RUN;**

# SELECCIÓN DE LA MUESTRA

- **PROC SURVEYSELECT DATA=UNIVERSO METHOD=SRS N=50  
OUT=MUESTRA\_ESTMAS OUTSIZE ;**
- STRATA ESTRATO;
- **RUN;**
  
- **PROC SURVEYMEANS DATA=MUESTRA\_ESTMAS TOTAL=NPOB SUM  
CVSUM;**
- BY ESTRATO;
- weight SamplingWeight;
- ratio YK/ ZK;
- VAR XK YK ZK;
- STRATA ESTRATO;
- ODS OUTPUT Statistics=TOTALES RATIO=RAZON;
- **RUN;**

# RESULTADOS

- ✓ Para cada muestra se calcula la diferencia entre las estimaciones del total y de la razón por estrato, así como el cociente entre los coeficientes de variación estimados.
- ✓ Se obtiene un promedio para cada tamaño de muestra evaluado.

## Resumen comparación de estimaciones

$n_h$	diff total	dif razón	cociente cve(total)	cociente cve(razón)
50	6.69E-12	-7.63E-19	1	1
100	8.73E-13	7.63E-19	1	1
150	2.24E-11	-2.08E-18	1	1
200	4.87E-12	-1.01E-18	1	1

*500 muestras para cada tamaño de muestra*

Conclusión: Resultados exactamente iguales.

# DISEÑO MAS<sup>2</sup>

- **SIMULACION DEL UNIVERSO DE ESTUDIO**
- **200 UPM de tamaño aleatorio entre 60 y 120.**
- **17832 individuos.**
- **data universo;**
- **do upm=1 to 200;**
- **NI=ROUND((90+60\*(ranuni(123)-0.5)));**
- **do ind=1 to NI;**
- **xk=2\*rangam(12345,100);**
- **ALEA1=RANUNI(23);**
- **ALEA2=RANUNI(45);**
- **YK=0;ZK=0;**
- **IF ALEA1<=0.9 THEN ZK=1;**
- **IF ALEA2<=0.1 THEN YK=1;**
- **IF YK=1 AND ZK=0 THEN YK=0;**
- **PEGA=1;**
- **output;**
- **end;**
- **end;**
- **run;**

# ESTIMACIONES CON S.A.S.

- `proc surveymeans data=muestra total=200 sum CVSUM;`
- `cluster upm;`
- `var xk YK ZK;`
- `ratio YK/ ZK;`
- `weight fexp;`
- `ODS OUTPUT Statistics=TOTALES RATIO=RAZON;`
- `run;`
- Se realizó el ejercicio de simulación para diferentes valores de las fracciones de muestreo en las dos etapas.

# RESULTADOS PARA TOTALES Y RAZONES

Como medida de comparación, se calcula la diferencia entre las estimaciones y se promedia para cada combinación de f1 y f2.

Diferencias estimación del total

f1	f2	
	0.2	0.4
0.2	1.0710E-09	-1.2480E-09
0.4	-6.4261E-10	8.3819E-11
0.6	1.5274E-09	-1.1642E-09

500 muestras en cada celda

Diferencias estimación de la

razón f1	f2	
	0.2	0.4
0.2	-2.69229E-17	-2.41196E-16
0.4	-1.92346E-16	6.21725E-17
0.6	-1.24595E-15	1.91236E-16

500 muestras en cada celda

Conclusión: Estimaciones de totales y razones exactamente iguales.

# RESULTADOS PARA COEFICIENTES DE VARIACION

Como medida de comparación, se calcula el cociente entre los errores de muestreo reales y los estimados por S.A.S y se promedia para cada combinación de f1 y f2.

Cociente entre cve para el total y porcentaje de varianza recogido en la primera etapa

f1	f2		f2	
	0.2	0.4	0.2	0.4
0.2	5.0003	5.0001	0.9999	0.99995
0.4	2.5016	2.5006	0.9987	0.9995
0.6	1.6721	1.6687	0.9935	0.9976

500 muestras en cada celda

Cociente entre cve para la razón y porcentaje de varianza recogido en la primera etapa

f1	f2		f2	
	0.2	0.4	0.2	0.4
0.2	5.0205	5.0163	0.9919	0.99350
0.4	2.6129	2.5830	0.9156	0.9368
0.6	2.0009	1.9456	0.6944	0.7343

500 muestras en cada celda

Conclusión: Coeficientes de variación estimados calculados con S.A.S son sustancialmente menores a los reales.

# COMPARACION DE LAS FORMULAS PARA LA ESTIMACION

Tradicional

$$\hat{V}_{MAS^2}(\hat{t}_{y\pi}) = \frac{N_1^2}{n_1} \left(1 - \frac{n_1}{N_1}\right) S_{\hat{t}_i, m_1}^2 + \frac{N_1}{n_1} \sum_i \frac{N_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right) S_{y, m_i}^2 = V_1 + V_2$$

En S.A.S

$$\hat{V}_{SAS}(\hat{t}_{y\pi}) = \frac{n_1(1-f_1)}{n_1-1} \sum_i (\hat{t}_i - \bar{t})^2 = n_1(1-f_1) S_{\hat{t}_i, m_1}^2$$

Luego,

$$V_1 = \left(\frac{N_1}{n_1}\right)^2 \hat{V}_{SAS}(\hat{t}_{y\pi})$$

S.A.S solo estima la varianza entre las unidades primarias de muestreo.

- Considerando que la primera etapa de muestreo recoja un porcentaje de variabilidad  $p$ .

$$V_1 = p \hat{V}_{MAS^2}(\hat{t}_{y\pi})$$

Se tiene que:

$$\hat{V}_{MAS^2}(\hat{t}_{y\pi}) \cong \frac{\left(\frac{N_1}{n_1}\right)^2 \hat{V}_{SAS}(\hat{t}_{y\pi})}{p} \qquad cve_{MAS^2}(\hat{t}_{y\pi}) \cong \frac{\left(\frac{N_1}{n_1}\right) cve_{SAS}(\hat{t}_{y\pi})}{\sqrt{p}}$$

Para el caso anterior, donde  $f_1=0.6$  y  $p=0.6944$ ;  $p=0.7343$

$$\frac{\left(\frac{N_1}{n_1}\right)}{\sqrt{p}} = \frac{1}{0.6\sqrt{0.6944}} = 2 \qquad \frac{\left(\frac{N_1}{n_1}\right)}{\sqrt{p}} = \frac{1}{0.6\sqrt{0.7343}} = 1.94498$$

# CONCLUSIONES

1. Para los diseños MAS y ESTMAS las estimaciones de totales, razones y coeficientes de variación son exactamente iguales.
2. Para el diseño MAS<sup>2</sup>, las estimaciones de totales y razones coinciden pero los coeficientes de variación no.
3. Es posible aproximar los errores de muestreo reales a partir de los estimados en SAS, conociendo la fracción de muestreo en la primera etapa y el porcentaje de variabilidad recogido en la misma.

**GRACIAS**