# MUESTREO EN POBLACIONES FINITAS 2016
## Fundamentos y Métodos

## Trabajo Práctico 2
## PARTE A

PALABRAS CLAVES: Muestreo por Conglomerados en dos etapas; Método de Sampford; Probabilidades de 1er y 2do orden en diseños en dos etapas.

## INTRODUCCIÓN

Cierta localidad consta de aproximadamente 22400 viviendas según el último censo de población y viviendas del año 2010 y están conglomeradas en 64 Radios $(M = 64)$ de aproximadamente 350 viviendas en promedio cada uno.

El Radio es una unidad de área que conglomera viviendas y surge como unión de otras unidades cartográficas-espaciales como la manzana (Mza), ver Fig. 1 y se emplea con fines censales, administrativos y muestrales. Para el muestreo es una unidad conveniente, por ejemplo define un área de trabajo para un encuestador y permite controlar la ineficiencia en términos de costos y organizativos de una muestra simple al azar de unidades elementales (viviendas) ya que ésta puede estar muy dispersa geográficamente.

Es así que en la etapa de su definición se desea mantener algunos criterios de regularidad con respecto al tamaño de los Radios, pretendiendo que no se modifiquen a través del tiempo. Vale recordar que tener en lo posible conglomerados de igual tamaño es un beneficio para cualquier diseño que los emplee, ya que no se introduce una variabilidad "extra" que impacta en el error muestral de las estimaciones. Por ejemplo, el efecto de diseño o *deff* al emplear conglomerados en vez de unidades elementales bajo un diseño muestral es una función del CV de los tamaños $N_i$ de los conglomerado[1] $i = 1,\ldots, M$

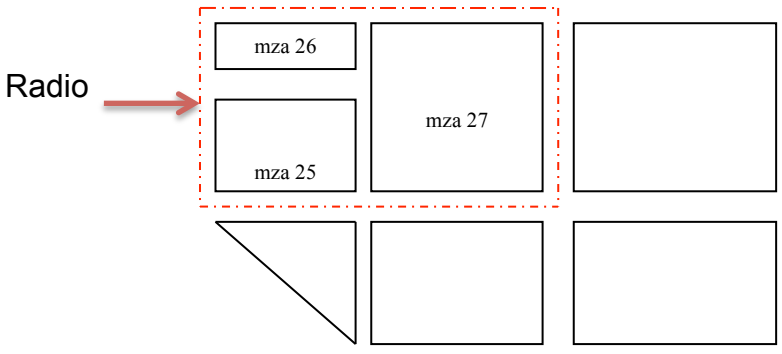


Fig. 1

Para confeccionar un diseño que emplee a los Radios como conglomerados en una primer etapa (de ahora en más Radio=Unidades Primarias de Muestreo=UPM) generalmente se busca vincularle la información recogida en un Censo previo permitiendo caracterizarlos. Esta estratificación de los Radios se logra a partir de una o más características

[1] Ver Chap 4, Remarks 4.2.1 a 4.2.3, en *Model Assisted Survey Sampling,* Sarndall-Wretman & Swensson.

poblacionales según información a nivel de viviendas (por ejemplo, densidad de viviendas por radio, tipo de vivienda, antigüedad, n$^{ro}$ de habitación, proporción de viviendas unipersonales, etc.) y/o a nivel de Población (por ejemplo, densidad de población por radio, proporción de población con el máximo o mínimo nivel de educación alcanzado por radio, proporción de niños menores a 10 años por radio, etc.).

Al estratificarlos se busca ganar precisión en las estimaciones finales, ya deterioradas por el efecto conglomerado señalado en párrafos anteriores. Y como es habitual entre los supuestos, se especula que la mayor parte de las características a estudiar por la encuesta tienen un alto o moderado grado de correlación con las intervinientes en la estratificación.

## CARACTERÍSTICAS DEL DISEÑO MUESTRAL DESEADO

Para la localidad que nos ocupa, se planea una encuesta de características socioeconómica y demográfica. Entre las necesidades a estimar están: la tasa de desempleo, el total de la población económicamente activa, estimaciones sobre el tipo de cobertura en salud que tiene la población, el ingreso medio de los hogares, etc.

Para alcanzar con el objetivo se propone un diseño muestral en dos etapas por conglomerados, estratificando a las UPM y seleccionándolas proporcional a un tamaño, a través del método de rechazo de **Sampford** definiendo así a la 1er etapa. La segunda etapa, consta de una selección de viviendas (Unidades Secundarias de Muestreo o USM) por **Muestreo Simple al Azar** en cada uno de las UPM seleccionadas de cada estrato.

Los conglomerados en cuestión ya fueron estratificados en 3 estratos en virtud de 4 características asociadas al nivel de educación de la población, un indicador de pobreza, el tipo y la antigüedad de las viviendas que componen al radio, todas disponibles gracias al censo. Como consecuencia los 64 radios se dividen en: 26 en estrato 1, 30 en el estrato 2 y 8 en el estrato 3. En el archivo "**radiosTP2.RData**" se listan todos los radios con un indicador del estrato (**Estrato**) de pertenencia, más la medida o tamaño a emplear para seleccionarlos, "**Tviv**" = Cantidad de Viviendas Ocupadas. La medida en cuestión surgió de un proceso de re-listado de las viviendas en terreno de cada uno de los 64 Radios que actualizó un listado surgido del Censo 2010, apuntando a reconocer altas, bajas o modificaciones en las viviendas de la localidad.

Para la determinación de la cantidad de Radios a seleccionar (*m)* en cada estrato y el tamaño de muestra final de viviendas en cada uno de los conglomerados ($n_i$), se minimizó la varianza del estimador de HT del total de desocupados en la localidad por estrato, sujeta a una función de costo esperado total fijo y dado por los recursos disponibles[2], para cada estrato.

[2] Para los que tienen interés en los lineamientos generales del procedimiento para determinar el tamaño de muestra de 1era y 2da etapa, ver apartado 12.8 "Allocation Problems in Two-Stage Sampling" del Chap 12 en *Model Assisted Survey Sampling*, Sarndall-Wretman & Swensson, en particular el Result 12.8.1, pág 473.

La resolución numérica de la minimización con datos Censales del año 2010[3] llevó a seleccionar **6** conglomerados en el estrato 1, **9** en el estrato 2 y **3** en el estrato restante. Como consecuencia la muestra final de primer etapa a **18** conglomerados de los 64.

Para la segunda etapa, la selección de viviendas dentro de los conglomerados seleccionados en la primera se decidió unificar los tamaños de muestra $n_i$ resultantes de la minimización en cada conglomerado según su estrato de pertenencia a **90**, **50** y **80** viviendas respectivamente. Se trató de respetar aproximadamente los tamaños surgidos de los cálculos, favoreciendo un aumento del tamaño especulando una probable no respuesta. Esto lleva a constituir una muestra final de **1230** viviendas para la localidad, distribuidas en **540** (6*90), **450** (9*50) y **240** (3*80) por estrato.

Se pide:

a) Seleccionar la muestra estratificada de 1er etapa bajo la selección proporcional a tamaño (**Tviv**) según el método de Sampford recurriendo al pakage "sampling" y empleando el archivo "**radiosTP2.RData**".
b) Presentar un detalle con las unidades seleccionadas y sus respectivas probabilidades $\pi_{hi}^{UP}$ y $\pi_{hij}^{UP}$ de primera etapa,
c) Determinar el vector de $\pi_{k/h}$, y la matriz de $\pi_{kl/h}$ para las viviendas en cada estrato,
d) ¿Qué valores toma $\Delta_{kl}$ para viviendas de distinto estrato?

*Observaciones:*

1) el subíndice $h$ hace referencia al estrato en cuestión al que pertenecen tanto los conglomerados como las viviendas, los subíndices $i, j$ a los conglomerados o UPMs y $k, l$ a las unidades elementales o viviendas.
2) retener la información de los puntos 1b)-1c) ya que serán centrales en la etapa de estimación del la parte 2 del TP.
3) tener presente el punto 1d) a la hora de componer las estimaciones de varianza o Cvs.

[3] El Censo indaga sobre la condición de actividad de la población, con lo cual para todos los cálculos se emplean los datos censales en cada Conglomerado o Radio Censal y en cada estrato.

# Trabajo Práctico 2
# PARTE B

PALABRAS CLAVES: Estimadores de Horvitz Thompson, de Hajek y por Razón; Estimaciones en Dominios; Estimación de la Varianza para Estimadores de Razón; Coeficiente de Variación; Componentes de la Varianza del Estimador; Efecto de Diseño.

La segunda parte del Trabajo Práctico 2 consiste en estimar totales, razones y/o tasas para la población y en algunos dominios con sus respectivas estimaciones de varianzas y/o CV y efectos de diseño. Para las estimaciones se empleará los datos que se encuentran en el archivo "**muestraTP2.RData**" que se asumirán el resultado de la encuesta de la muestra seleccionada en la 1ra Parte del TP2.

Recordar que la muestra 1230 viviendas era en dos etapas con selección de radios proporcional al tamaño de viviendas ocupadas según Censo 2010 (previamente estratificados) y un submuestreo de tamaño fijo según los estratos a través de un MSA en cada radio seleccionado de 1er etapa.

El archivo "**muestraTP2.Rdata**" está organizado a nivel de viviendas[4] y contiene las siguientes características e información:

**Estrato** (estratificación de las unidades de 1er etapa),
**UPM** (unidades de 1er etapa o UPM),
**Nroviv** (número de la vivienda en la UPM),
**totper** (total de personas en la vivienda encuestada),
**hog10** (**0**=hogar sin menores de 10 años, **1**=hogar con menores de 10 años),
**SaludF** (total de mujeres del hogar con cobertura de Salud),
**SaludM** (total de hombres del hogar con cobertura de Salud),
**Salud55** (total de personas en el hogar mayor de 55 años con cobertura de Salud),
**Ftedad2 (**total de miembros del hogar que están en la **PEA**[5] y 15<=edad<=29)
**Ftedad3 (**total de miembros del hogar que están en la **PEA** y 30<=edad<=49)
**Ftedad4 (**total de miembros del hogar que están en la **PEA** y 50<=edad<=65)
**Oedad2 (**total de miembros del hogar Ocupados y 15<=edad<=29)
**Oedad3 (**total de miembros del hogar Ocupados y 30<=edad<=49)
**Oedad4 (**total de miembros del hogar Ocupados y 50<=edad<=65)
**IngresoH** (Ingreso del hogar sin subsidios o planes)

*Observaciónes importantes antes de comenzar con las estimaciones*:

1) la variable **UPM** tiene una numeración correlativa de 1 a 6, de 1 a 9 y de 1 a 3 según el estrato de pertenencia y relaciona a la UPM del archivo "**MuestraTP2.Rdata**" con la selección de Radios alcanzada por "sampling" sobre "**radiosTP2.RData**".
Por ejemplo para el estrato 1, **UPM=**(1,2,3,4,5,6) relaciona en forma natural a cualquier selección de 6 radios de 1er etapa, y si la selección por Sampford fue

[4] Se considera una vivienda = un hogar

[5] PEA = Persona Económicamente Activa y considerada en condiciones para formar parte de la fuerza de trabajo. Sobre ésta subpoblación se realizan los cálculos asociados a los ocupados y/o desocupados. Menores de 15 años y mayores de 65 años son excluidos de la PEA.

**Radio=** (5,28,29,33,34,35), UPM=1 es equivalente Radio=5 y UPM=2 equivale a Radio=28 y así sucesivamente.

Por otro lado como en cada una de las UPMs seleccionadas en éste estrato tuvo una submuestra de 90 viviendas como consecuencia en cada **UPM** y en ése estrato, en el archivo "**MuestraTP2.Rdata**" el valor que toma **Nroviv** va de 1 a 90 en forma consecutiva por UPM. Esta modalidad se repite para el estrato 2, donde **Nroviv** va de 1 a 50 en cada **UPM** seleccionada y para el estrato 3, en cuyo caso **Nroviv** va de 1 a 80 dado que 80 es el total de viviendas seleccionadas por **UPM** en dicho estrato.

2) Como consecuencia del punto 1) el archivo no cuenta con las probabilidades de inclusión $\pi_{hi}^{UP}, \pi_{k/hi}, \pi_{k/h}$, Por lo tanto deberán ser tenidas en cuenta e imputadas a los registros del archivo "**MuestraTP2.Rdata**" según las que correspondan a su selección para poder lograr las estimaciones que se piden.

3) Lo señalado en el punto 2) también aplica a las probabilidades de 2do orden del diseño ya que dependen de la muestra seleccionada en 1er etapa en cada estrato y del submuestreo en la 2da. Es muy importante recordar que las las matrices de 2do orden pueden ser calculadas para cada estrato por separado y que existe independencia en la selección de unidades en cada uno de ellos, virtud que simplifica las estimaciones de la varianza o error muestral.

Se pide a partir de los datos del archivo estimar (en todos los casos se deberá consignar la estimación del parámetro, el CV y el deff):

a) el tamaño de la población de la localidad
b) la proporción de población con cobertura de salud y cobertura según sexo,
c) la proporción de la población mayor a 55 años con cobertura de salud,
d) el total de la PEA, el total de desocupados y la tasa de desocupación (desocupados/PEA) para la localidad en cuestión,
e) el Ingreso promedio en los hogares,
f) el ingreso promedio en los hogares con menores de 10 años,
g) el ingreso promedio en los hogares por tamaño (según clasificación: **A=**hogar unipersonal, **B=**hogar con 2 personas, **C=**hogar con 3 o 4 personas, **D=**hogar con 5 o más personas)