# Diseño Muestral de las Encuestas dirigidas a la Población en el INE

Febrero,2014



# ¿ Por qué son importantes las estadísticas ?



La Ley atribuye al INE un papel destacado en la actividad estadística pública y le encomienda:

- La **realización de las operaciones estadísticas** de interés nacional: (censos demográficos y económicos, cuentas nacionales, estadísticas demográficas y sociales, indicadores económicos y sociales, coordinación y mantenimiento de los directorios de empresas, formación del Censo Electoral...).

- La **formulación del Proyecto del Plan Estadístico Nacional** con la colaboración de los Departamentos Ministeriales y del Banco de España; la propuesta de normas comunes sobre conceptos, unidades estadísticas, clasificaciones y códigos, etc.

- Las **relaciones en materia estadística** con las CCAA y los Organismos Internacionales especializados y, en particular, con la Oficina de Estadística de la Unión Europea (EUROSTAT).

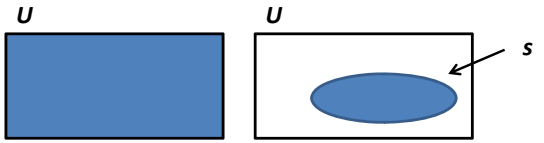


En la ejecución de su actividad estadística el INE realiza, *entre otras*, dos grandes tipos de operaciones estadísticas: **Censos y Encuestas por Muestreo**

Dada una población que denotamos por ***U,*** hablamos de :

- **Censo** : cuando se realiza un estudio exhaustivo de la población ***U.***

- **Encuesta por Muestreo**: cuando observamos un subconjunto (***s***), para inferir datos del conjunto de la población ***U.***



## *Censo vs Encuestas*

### Censo

✓Desagregación Geográfica de los resultados

✓Muy costoso

✓Sin errores de muestreo

✓Errores ajenos al muestreo

### Encuestas

✓Más económicas y frecuentes

✓Con menos errores ajenos al muestreo

✓Resultados Rápidos

✓Menos limitaciones en las características a investigar.

**Importancia**

**Proporcionan a las investigaciones por muestreo información necesaria en:**

- **Preparación de las Bases de Muestreo (MARCOS)**
- **Procesos de ESTRATIFICACIÓN**
- **Procesos de ESTIMACIÓN**



## PRINCIPALES ETAPAS DE UNA ENCUESTA POR MUESTREO :

**Determinación de objetivos**

↓

**Diseño de la muestra**

↓

**Trabajos de Campo**

↓

**Tratamiento de la Información**

↓

**Difusión de Datos**



## Introducción al Diseño Muestral

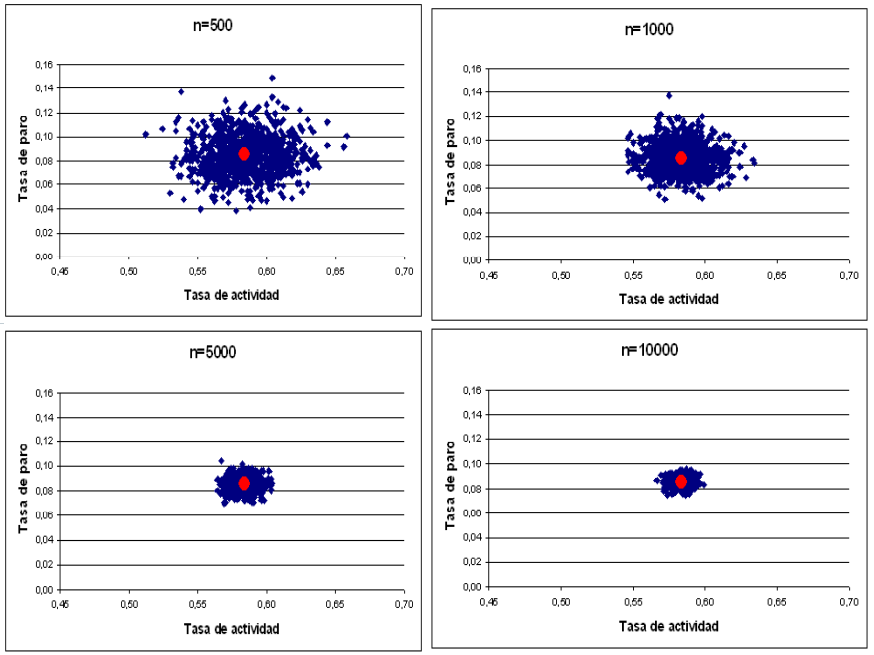
**Objetivo :** Obtener *información* precisa sobre un *parámetro poblacional* a partir de una *muestra* de unidades elementales.

**Información previa necesaria**

- Variable objetivo
- Tipo de estimaciones
- Precisión de las estimaciones
- Tablas
- Experiencias anteriores



## *Lorca 2001. Simulación (m.a.s. 1000 iteraciones)*



# Tipos de Muestreo más frecuentemente utilizados

Se consideran dos grandes grupos:

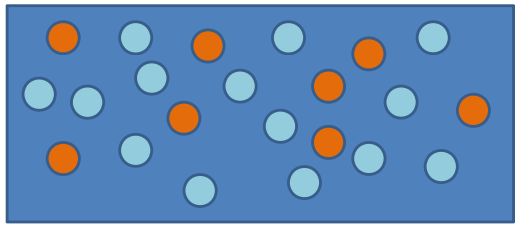
- ***Muestreo probabilístico:*** Todas las unidades del marco tienen probabilidad conocida y mayor que cero de pertenecer a la muestra
- ***Muestreo no probabilístico:*** Se desconocen las probabilidades de pertenencia a la muestra

***El Instituto Nacional de Estadística(I.N.E.) sólo utiliza tipos de muestreo probabilísticos, ya que permiten obtener una medida de la precisión de las estimaciones***



## Muestreo Aleatorio Simple

Selección de unidades elementales con probabilidades iguales. Dada una población de tamaño N, se selecciona una muestra formada cualesquiera n elementos de la población.



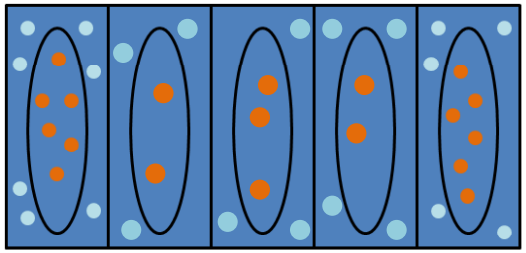


## Muestreo Estratificado

Sea una población heterogéa de N unidades $\{U_i\}$ . Ésta se subdivide en L **subpoblaciones** lo más **homogéneas** posibles y no solapadas, llamadas **estratos** de tamaños $N_1,\ldots, N_L$ .

La muestra estratificada de tamaño *n* se obtiene seleccionando $n_h$ elementos (h=1,…,L) de cada uno de los L estratos en los que se subdivide la población de forma independiente.

**L: nº de estratos**
**$N_h$ : tamaño poblacional del estrato h con h=1,…,L** → **$N=\sum N_h$**
**$n_h$ : tamaño muestral del estrato h=1,…, L** → **$n=\sum n_h$**



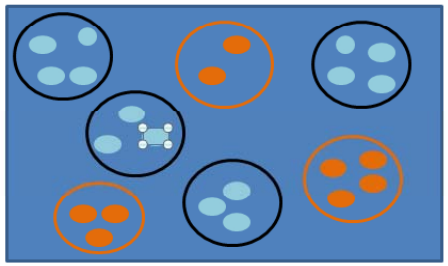


## Muestreo por Conglomerados sin submuestreo

Sea una población finita constituida por **N** unidades elementales, que de alguna manera natural o artificial, están agrupadas en **M** unidades mayores que se denominan **conglomerados**.

Este tipo de muestreo, consiste en la selección de muestras constituidas por todas las unidades elementales de los **m** conglomerados elegidos.

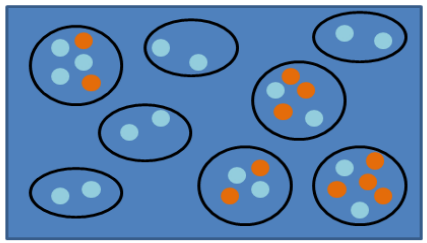
**Ejemplo : N=23, M=7, m=3**



## Muestreo por Conglomerados con submuestreo

Supongamos que partimos de una población similar a la considerada en el muestreo de conglomerados sin submuestreo y que las unidades elementales de los conglomerados fuesen parecidas entre si. En este caso, tal vez un pequeño nº de ellas constituirían una muestra representativa sin necesidad de utilizar todo el conglomerado muestral.

Lo primero es seleccionar una muestra de m conglomerados .En segunda etapa, se selecciona una muestra de unidades elementales en cada unidad primaria elegida.



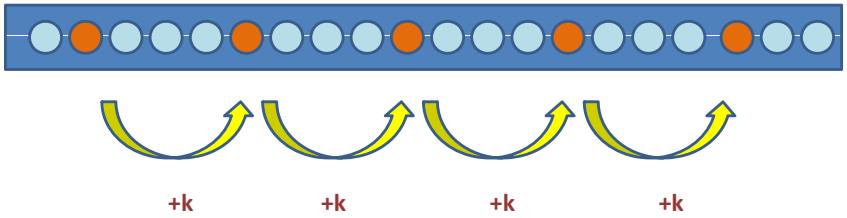


## Muestreo Sistemático

Consideramos una población de tamaño N y agrupamos sus elementos es n zonas de tamaño k, donde k=N/n. Para extraer una muestra de tamaño n, se elije un nº aleatorio j € [1,k] y de manera que la muestra quedad formada por las unidades $u_j, u_{j+k}, \ldots u_{j+(k-1)k}$.

Luego la selección de la primera unidad determina toda la muestra. Según esto para una nuemeración dadade las unidades de la población, el espacio muestral está constituido por k posibles muestras , y todas ellas con la misma pobrabilidad de ser seleccionadas (1/k)

**Ejemplo : N=20 , n=5, k=4, j=2**






**Muestreo por cuotas:**

Selección de una muestra equilibrada según variables conocidas, por ejemplo, por grupos de edad y sexo(cuotas).

**Desventaja:**

Es frecuente utilizar este tipo de muestreo cuando no se dispone de marco. Por ejemplo, seleccionar la muestra a la salida del metro, de forma opinática, hasta que se tenga el tamaño muestral necesario en cada cuota. Sería entonces un muestreo no probabilístico y por tanto con estimaciones de sesgo y precisión desconocidos

## Muestreo por rutas aleatorias:

Normalmente asociado al muestreo por cuotas, realiza la selección de la muestra mediante itinerarios aleatorios predefinidos.

**Ventajas:**

- No es precisa la elaboración de un marco
- Reducido coste de los trabajos de campo

**Desventajas:**

Falta de control adecuado de las probabilidades de selección



# El diseño de la muestra
## Esquema

1. ÁMBITO
2. MARCO DE LA ENCUESTA
3. TIPO DE MUESTREO
4. CRITERIOS DE ESTRATIFICACIÓN
5. TAMAÑO Y AFIJACIÓN DE LA MUESTRA
6. SELECCIÓN
7. DISTRIBUCIÓN DE LA MUESTRA EN EL TIEMPO
8. RENOVACIÓN PARCIAL DE LA MUESTRA
9. ESTIMADORES *(TÉCNICAS DE CALIBRADO)*
10. ERRORES DE MUESTREO
11. ACTUALIZACIÓN DE LAS UNIDADES DE MUESTREO

***Nota:Las siguientes características son comunes a las encuestas de hogares(con excepciones como la Encuesta de Morbilidad Hospitalaria, Personas sin hogar, etc..)***

**1.Ámbito:** se contempla desde una triple óptica

**Poblacional:** Población que reside en viviendas familiares principales

**Temporal**

**Geográfico**

**2.Marco:**

**-Marco de Unidades Primarias:** Relación de secciones censales

-**Marco de Unidades Secundarias**: Relación de viviendas familiares principales

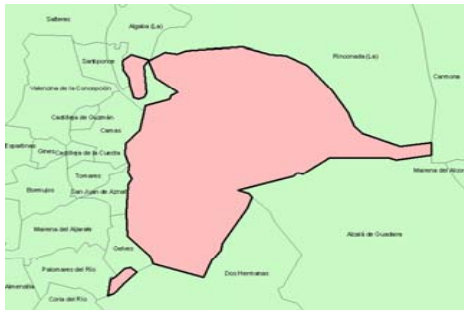
Ambos marcos se elaboran, bien con información del último **Censo**, bien con información del **Padrón Continuo**

**3.Tipo de muestreo:** Bietápico con estratificación de unidades de primera etapa.

-**Unidades primarias:** Secciones censales

-**Unidades secundarias:** Viviendas familiares principales







**4.Criterios de Estratificación**

La Estratificación en la Encuestas a Hogares se realiza sobre las unidades de primera etapa*( secciones censales)*

**-Criterio geográfico(Estratos):** Según el tamaño del municipio al que pertenecen

Ejemplo EPA:

**Municipios Autorrepresentados: Estratos 1-2-3**

**Municipios *Correpresentados*** :

| Estratos | Población |
|---|---|
| 4 | 50.000 - 100.000 |
| 5 | 20.000 - 50.000 |
| 6 | 10.000 - 20.000 |
| 7 | 5.000 - 10.000 |
| 8 | 2.000 - 5.000 |
| 9 | < 2.000 |

**-Criterio socioeconómico (Subestratos):**establecido a partir de información censal. Las definiciones de los subestratos se establecen aplicando técnicas de análisis de conglomerados.



**5.Tamaño de la muestra**

Para su cálculo se toma en consideración:

1- La desagregación requerida para las estimaciones

2- La dispersión de la(s) variables(s) objetivo('σ')

3- Precisión y fiabilidad establecidas por el Servicio Promotor

Resulta muy útil la experiencia de otras encuestas anteriores

De acuerdo con lo anterior se establece:

- Un número de secciones muestrales por estrato
- Un número constante de viviendas por sección

## 6.Selección de la muestra

•**Unidades primarias(secciones censales):** Se eligen con probabilidad proporcional al tamaño.

•**Unidades secundarias(viviendas):** Selección con probabilidad igual(muestreo sistemático).

De esta forma se obtienen **muestras autoponderadas** en cada estrato

$$P\left(V_{ijh}\right)=P\left(S_{jh}\right).P\left(\frac{V_{ijh}}{S_{jh}}\right)=K_h\cdot\frac{V_{jh}}{V_h}\cdot\frac{m}{V_{jh}}=\frac{K_h\,.\,m}{V_h}$$



## 7.Distribución de la muestra en el tiempo

Se procura distribuir la muestra de forma uniforme a lo largo del ámbito temporal en el que se desarrolla.

Para ello las variables que, normalmente, se toman en consideración son:

- Semana
- Provincia o Comunidad autónoma
- Estrato

Ejemplo:: Distribución por Estrato y Turno de Rotación de la EPA

| | | TR | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CPRO | ESTRATO | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| 11 | 1 | 2 | 2 | 2 | 2 | 3 | 2 | 13 |
| | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 13 |
| | 3 | 1 | 1 | 1 | 1 | . | 2 | 6 |
| | 4 | 4 | 5 | 5 | 4 | 4 | 4 | 26 |
| | 5 | 1 | 1 | 1 | 2 | 1 | 1 | 7 |
| | 6 | . | 1 | 1 | 1 | 2 | 1 | 6 |
| | 7 | 2 | 1 | 1 | 1 | 1 | 1 | 7 |
| | Total | 13 | 13 | 13 | 13 | 13 | 13 | 78 |
| 51 | 1 | 2 | 3 | 2 | 2 | 2 | 2 | 13 |
| | Total | 2 | 3 | 2 | 2 | 2 | 2 | 13 |



## 9.Estimadores

Pasos para la obtención de estimadores:

- **Estimador insesgado de expansión(Horvitz-Thompson):** Compensa las desiguales probabilidades de selección.
- **Corrección de la falta de respuesta:** Corrige el sesgo producido en las estimaciones por la falta de respuesta total de algunos elementos.
- **Calibrado con fuentes externas:**Reduce la varianza de las estimaciones mediante la utilización de fuentes auxiliares externas y puede actualizar la estimación en el tiempo.

Como resultado de este proceso se obtiene finalmente un factor de elevación para cada elemento de la muestra efectiva.



## 9.Estimadores

### 9.1. Estimador insesgado de expansión(H-T)

Recordamos que la probabilidad de pertenecer a la muestra de una vivienda 'i' de la sección 'j' del estrato 'h' viene dada por:

$$P\left(V_{ijh}\right)=P\left(Sec_{jh}\right).P\left(V_{ijh}/Sec_{jh}\right)=K_h\cdot\frac{V_{jh}}{V_h}\cdot\frac{m}{V_{jh}}=\frac{K_h\,.\,m}{V_h}$$

Donde :
Kh son las secciones de la muestra en el estrato "h"
m es el número de viviendas muestrales por sección

Según lo anterior, la probabilidad de pertenecer a la muestra se puede expresar por:

$$P\left(V_{ijh}\right)=\frac{v_h^t}{V_h}$$

Siendo vth el número teórico de viviendas de la muestra en el estrato "h".

Por tanto el estimador H-T tendrá la expresión:

$$\hat{Y}_{H-T}=\sum_h\frac{V_h}{v_h^t}\cdot\sum_{i\in h}y_i$$

## 9.Estimadores

### 9.2. Corrección de la Falta de Respuesta

La probabilidad de respuesta por estrato la podemos estimar por:

$$P_{Rh} = \frac{v_h}{v_h^t}$$

Donde vh representa la muestra efectiva de viviendas en el estrato h.

Por tanto el estimador corregido será:

$$\hat{Y}_{H-TCorr} = \sum_h \frac{V_h}{v_h^t} \cdot \frac{v_h^t}{v_h} \sum_{i \in h} y_i = \sum_h \frac{V_h}{v_h} \sum_{i \in h} y_i = \sum_h \hat{Y}_{H-TCorr}(h)$$



## 9.Estimadores

### 9.2. Técnicas de Calibrado

Un estimador está calibrado respecto de una o varias variables auxiliares, si las estimaciones de los totales poblacionales de estas variables, coinciden con los valores poblacionales conocidos

***Idea intuitiva 1:*** *"Si se estima correctamente la variable auxiliar, lo mismo ocurrirá con la variable objetivo"*

***Idea intuitiva 2:*** *"Si no se estima bien la sencilla variable auxiliar, a saber qué pasará con la variable objetivo"*

***Hay muchas formas de calibrar las estimaciones. Habrá que buscar cuál es la más razonable.***



## 9.Estimadores

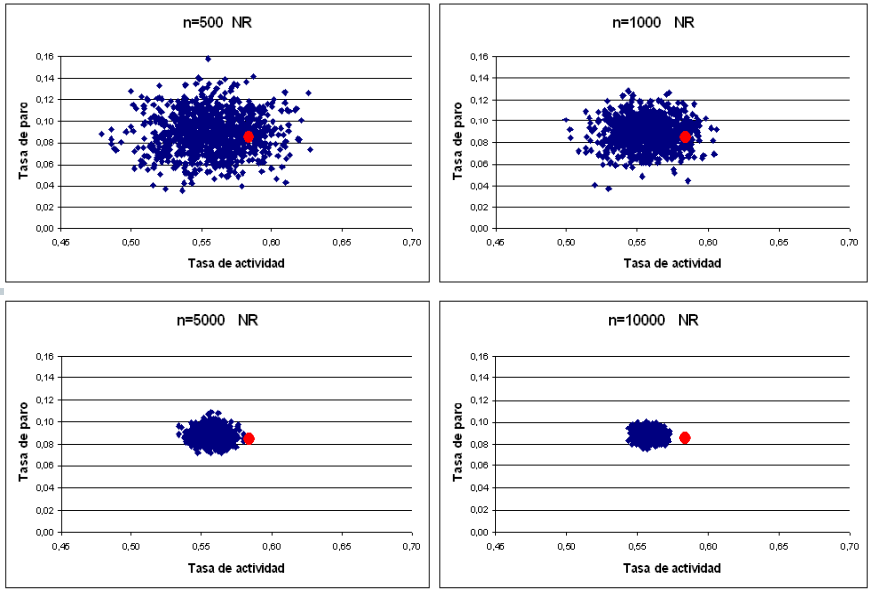
### 9.2. Técnicas de Calibrado. Objetivos

El objetivo de las técnicas de calibrado es utilizar **información auxiliar** disponible, de forma que mejore la **calidad de las estimaciones** en los siguientes aspectos:

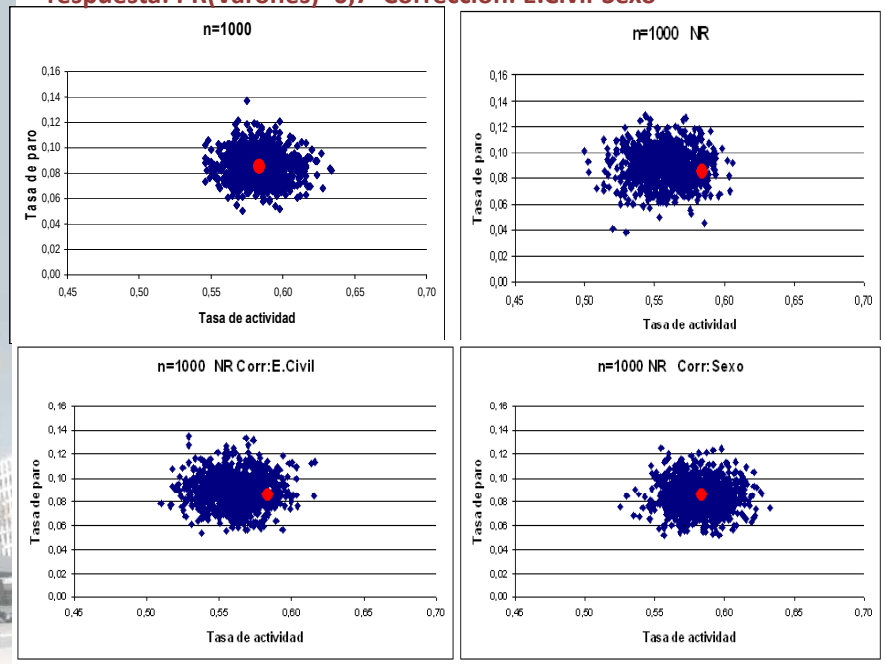
- Incremento de la precisión
- Disminución del sesgo producido por la falta de respuesta
- Aumento de la coherencia en la producción estadística:
  - Estimaciones comunes en encuestas diferentes
  - Estimaciones procedentes de submuestras
  - Etc, etc, etc...



***Lorca 2001. Simulación (m.a.s. 1000 iteraciones) No respuesta: PR(Varones)=0,7***



## Lorca 2001. Simulación (m.a.s. 1000 iteraciones n=1000) No respuesta: PR(Varones)=0,7 Corrección: E.Civil-Sexo






## 9.3. Técnicas de Calibrado. Estimador calibrado

**Partiendo del estimador** $\hat{Y} = \sum_{k \in s} d_k \, y_k$

**Se pretende encontrar otro** $\hat{Y}_{CAL} = \sum_{k \in s} w_k \, y_k$

**Que verifique:**

- **Estimar correctamente los totales poblacionales de las variables auxiliares (X).Cumpla las " condiciones de equilibrio"**

$$\sum_{k \in s} w_k \, x_k = X$$

- **Los nuevos pesos $\{w_k\}$ se diferencien lo menos posible de los pesos del estimador de diseño $\{d_k\}$**



## 9.3. Técnicas de Calibrado. Planteamiento del Problema

**Planteamiento del problema:**

**1- Elección de una función de distancia G, de argumento :**

$x = \frac{w_k}{d_k}$ , **positiva y convexa, y tal que** $\begin{cases} G(1) = G'(1) = 0 \\ G''(1) = 1 \end{cases}$

**2- Cálculo de los $w_k$ solución del problema:**

$$\min_{w_k} \sum_{k \in s} d_k \, G\left(\frac{w_k}{d_k}\right)$$ **con la condición** $$\sum_{k \in s} w_k \, x_k = X$$

**donde,** $x_k' = (x_{k1}, x_{k2}, \ldots, x_{kJ})$ $X' = (X_1, X_2, \ldots X_J)$



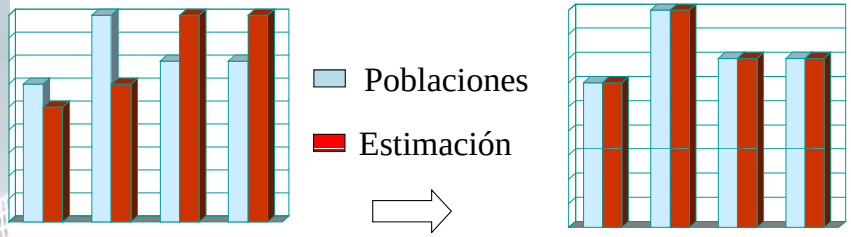
## 9.3. Técnicas de Calibrado. Aplicación : CALMAR

Para la resolución práctica de este problema se ha utilizado el software CALMAR (**CAL**age sur **MAR**ges) programado por el INSEE (Institut National de la Statistique et des Études Économiques) de Francia

- CALMAR es una macro pública de SAS y fácil de usar. El INSEE facilita documentación sobre su utilización, y hasta la fecha lo ha adaptado a las sucesivas versiones de SAS.
- Disponible en: www.insee.fr ( Définitions et Méthodes Outils Statistiques)
- Desarrollo informático: Olivier Sautory(INSEE)
- Teoría: Särndal, Deville y Sautory("Generalized Raking Procedures in Survey Sampling" JASA 1993 Vol.88, No.423)
- Permite la utilización de cuatro funciones de distancia, dos de las cuales acotan la variación de los factores calibrados.
- Se comprueba experimentalmente que las cuatro funciones de distancia de CALMAR proporcionan estimaciones similares con muestras suficientemente grandes.

Las variables auxiliares que se emplean en las Encuestas a Hogares en el INE son, en general:

1- Grupos de edad y sexo
2- Tamaños de los hogares
3- Nacionalidad






## 10.Errores de Muestreo . Definición

El error **absoluto** de muestreo se define como la **raíz cuadrada de la varianza del estimador.**

$$\varepsilon_a = \sqrt{V(\hat{X})} = \sigma(\hat{X})$$

El error **relativo** de muestreo (Coeficiente de Variación) se define como la relación entre el error absoluto y la estimación.

$$\varepsilon_r = \frac{\sqrt{V(\hat{X})}}{\hat{X}} = \frac{\sigma(\hat{X})}{\hat{X}}$$



## 10.Errores de Muestreo . Utilidad

El conocimiento de la varianza de un estimador permite:

- Al **usuario** conocer el grado de fiabilidad de los datos.
- Al **diseñador** tomar decisiones entre diseños alternativos

El estadístico, al publicar el error de muestreo proporciona al usuario un intervalo numérico que presenta una cierta confianza, medida en términos de probabilidad, de contener el valor verdadero que se desea estimar.



## 10.Errores de Muestreo . Cálculo

- **Procedimientos directos:** Utilización de la fórmula de la varianza, de acuerdo al diseño de la encuesta. Se suelen utilizar en encuestas de tipo económico.

- **Procedimientos Indirectos**: Para diseños complejos se han diseñado métodos que permiten utilizar fórmulas sencillas. Se utilizan en las encuestas de hogares.

Los más utilizados son:

- **Basados en replicaciones del diseño**
  - oMétodo de los grupos aleatorios
  - o " de los conglomerados últimos.
- **Basados en replicaciones de la muestra**
  - o " de las semimuestras reiteradas.
  - o " Jackknife
  - o " Bootstrap

## 11. ACTUALIZACIÓN DE LAS UNIDADES DE MUESTREO

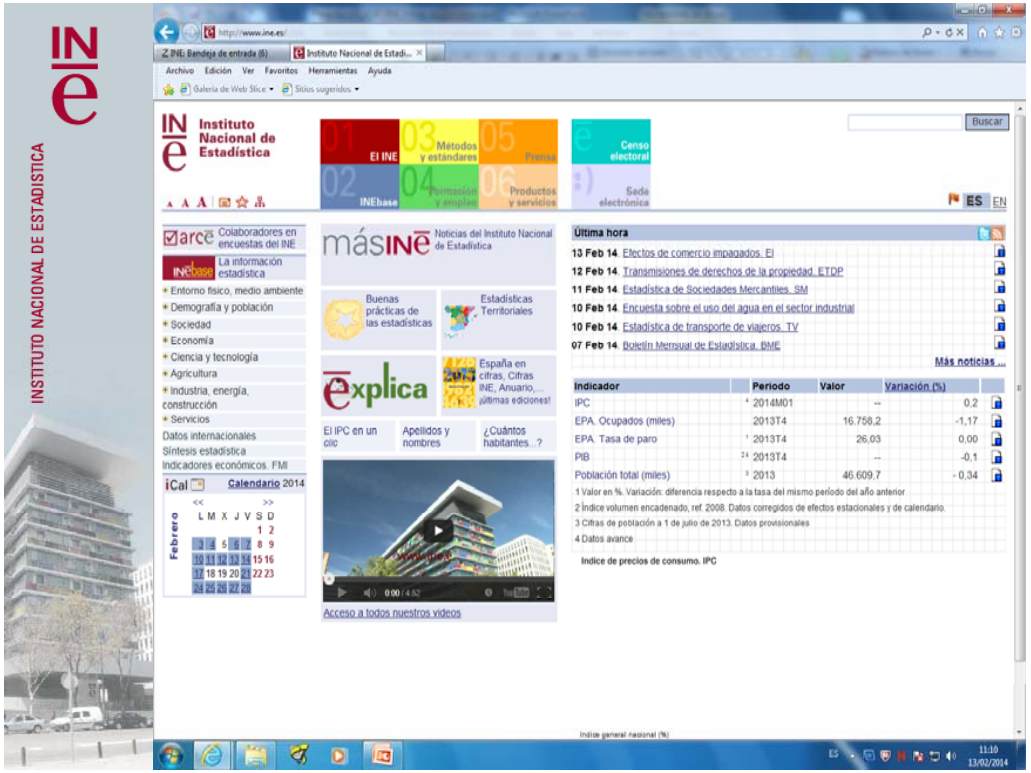
Las continuas variaciones de población ya sea en sus características, o bien en su distribución espacial exigen realizar actualizaciones en los marcos, que repercuten en la estructura muestral.

En líneas generales hay que considerar tres tipos de actualizaciones:

**Actualización en el marco de viviendas,** restringida a las secciones de la muestra.

**Actualización de secciones censales** como consecuencia de variaciones (particiones, fusiones,….) en las unidades primarias seleccionadas para la muestra.

**Actualización de carácter general** relativa a todas las secciones y viviendas, cuando se realizan los Censos de Población.




# Agradecimientos