

# **BASES DE DATOS FISCALES SISTEMAS DE INFORMACIÓN PARA LA MICROSIMULACIÓN**

## **Fuentes de datos del IEF**

**César Pérez López**

**Instituto de Estudios Fiscales  
Universidad Complutense de Madrid**

# Modelos de microsimulación

## Concepto

Son programas informáticos que partiendo de muestras representativas de la población, reproducen la estructura de políticas fiscales, ya sean de ingreso y/o de gasto, y simulan los efectos que una determinada reforma tendría en la población real.

## Principal ventaja

El uso de muestras representativas de la población y la cantidad de datos que proporcionan permite a los modelos de microsimulación comparar políticas alternativas de ingreso y gasto público con gran precisión en términos de recaudación, equidad y eficiencia.

# Modelos de microsimulación

## Tipología

**Modelos Tax-Benefit:** simulan simultáneamente políticas de ingreso y gasto público.

**Modelos particulares:** simulan una política específica de ingreso o gasto público.

**Modelos estáticos:** los datos de la muestra permanecen invariables.

**Modelos dinámicos:** los datos de la muestra se actualizan. Puede ser una actualización monetaria (PIB, IPC...) y/o demográfica.

**Modelos sin comportamiento:** no contemplan la reacción de los individuos ante los cambios.

**Modelos con comportamiento:** predicen reacciones de los individuos (oferta de trabajo, ahorro...)

# Modelos de microsimulación

## Modelos de microsimulación disponibles en España

**SINDIEF** (Sanz, Castañer, Romero, Fernández, 2001). Modelo particular de impuestos indirectos, estático, con comportamiento de la demanda.

**SIRPIEF** (Sanz, Castañer, Romero, Prieto, Fernández, 2004) Modelo particular de IRPF, estático, con comportamiento de la oferta de trabajo.

**ESPASIM** (Mercader, Levy, Planas, 2002). Modelo tipo tax-benefit, estático, sin comportamiento. Escenario anterior a Ley 40/1998 (reforma del 99)

**SIMBBVA/Gladhispania** (FBBVA/Spadaro y Oliver, 2005). Modelo particular de IRPF y cotizaciones sociales, estático, sin comportamiento: Escenario anterior a Ley 46/2002 (reforma del 2003)

# **Modelos de microsimulación**

## **Simuladores actuales del IEF**

### **SIMULADORES DE IMPOSICIÓN DIRECTA:**

**Impuesto sobre la renta de las personas Físicas (IRPF).**

**Impuesto de sociedades.**

**Impuesto sobre el Patrimonio.**

### **SIMULADORES DE IMPOSICIÓN INDIRECTA:**

**Impuesto sobre el valor añadido**


**Impuestos Especiales.**

### **SIMULADORES TAX-BENEFICT: EUROMOD**

**Simulador de impuestos y prestaciones para la Unión**

**Europea** que permite calcular, de manera comparable, los efectos sobre las rentas familiares y sobre los incentivos al empleo de cambios en los impuestos, en el diseño o cuantía de las cotizaciones o de algunas prestaciones sociales para la población de cada país y para la UE en conjunto.

### **SIMULADORES DE GASTO: PENSIONES**



# **HERRAMIENTAS DE SIMULACIÓN IMPOSITIVA DEL INSTITUTO DE ESTUDIOS FISCALES**

*Autor: Instituto de Estudios Fiscales<sup>(\*)</sup>*

**DOC. n.º 16/2011**

# LOS DATOS

- TRANSPARENCIA FISCAL Y MARCO
- FUENTES DE DATOS Y TRABAJOS
- MUESTRAS IRPF: DISEÑO
- REPRESENTATIVIDAD Y LIMITACIÓN
- NO OBLIGADOS NO DECLARANTES
- PANELES DE DATOS: DISEÑO
- BADESPE

# TRANSPARENCIA FISCAL

**PROGRAMA DE  
TRANSPARENCIA FISCAL** →

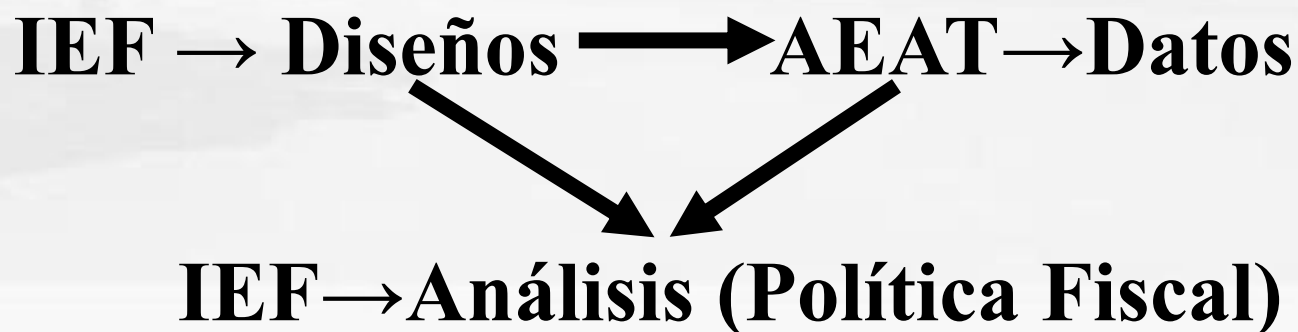
**-Elaboración y difusión BASES DE MICRODATOS FISCALES:**

- . Paneles de declarantes**
- . Muestras anuales**
- . Muestras no-declarantes**

**-BADESPE: Base de Datos del Sector Público Español**

# MARCO DE REFERENCIA

**Convenio entre la AEAT y el IEF  
Para el suministro de información  
Con fines estadísticos**



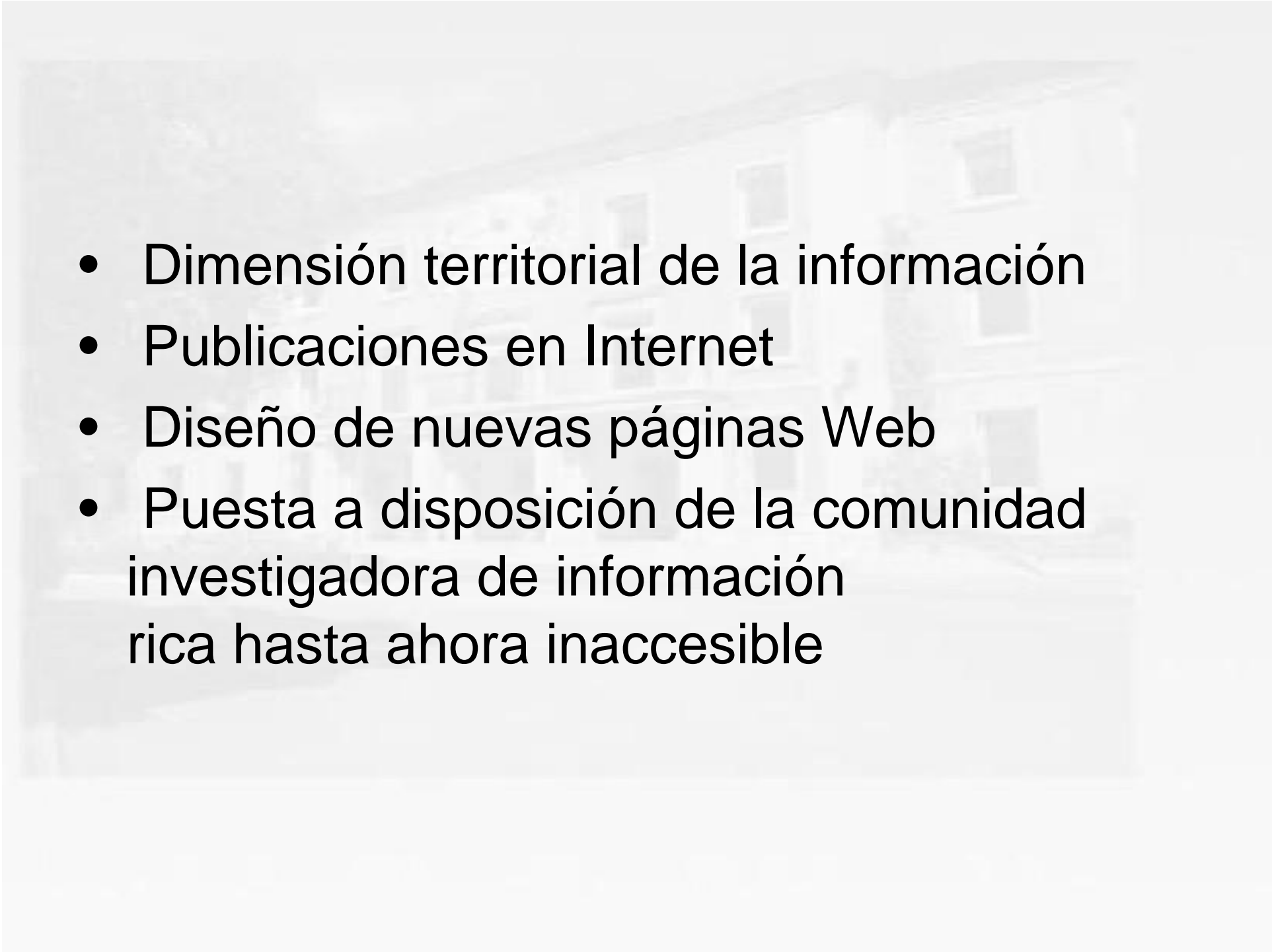
## **PLAN DE TRANSPARENCIA**

- Acuerdo de la Comisión Delegada del Gobierno para Asuntos Económicos sobre **mejoras en la transparencia en los ámbitos estadístico y económico**
- Espaldarazo a la **puesta a disposición de nuevas fuentes de información**
- Estado, CC AA y Entes Locales velarán por el suministro de **información estadística en tiempo y forma y sin restricciones**

- **Instaurar de forma definitiva y duradera** el plan de transparencia
- **Principios de transparencia:** Toda la información, buena difusión, disponible pronto y útil para el análisis y estudio
- Competencias plenarias del Estado para regular y ejecutar estadísticas con fines estatales
- Publicaciones **periódicas** con calendario de **información útil, bien desagregada, en soportes accesibles** y con técnicas modernas

## **LOGROS DEL PLAN DE TRANSPARENCIA**

- **Difusión normalizada, periódica y con calendario explícito.**
- Mejoras en la información disponible
- Implicación de todos los Ministerios:
- En Economía y Hacienda se involucran AEAT, IEF, IGAE, etc.

- 
- Dimensión territorial de la información
  - Publicaciones en Internet
  - Diseño de nuevas páginas Web
  - Puesta a disposición de la comunidad investigadora de información rica hasta ahora inaccesible

## ¿QUÉ FALTA?

- Recopilación de determinada información a través de nuevas encuestas
- Enlaces históricos de series
- Publicación de metodologías extensas
- Desarrollo de páginas Web adecuadas
- Acceso a información dinámica en la Web
- Mejorar la difusión
- Publicidad adecuada sobre la información disponible
- Contrastes de la calidad de la información

## **ESFUERZO DEL IEF - MICRODATOS**

- **El IEF adquirió el compromiso de puesta a disposición pública de muestras de microdatos procedentes de declaraciones tributarias**
- **Muestras de declarantes y no obligados no declarantes de IRPF 2002 a 2009**

- **Panel de declarantes de IRPF con información histórica 1982/1998**
- **Panel de declarantes de IRPF con información histórica 1999/2011**
- **Todas las muestras y paneles contienen información desagregada por criterios territoriales y constituyen un valor añadido a la disponibilidad de información territorializada (estratos)**

# **FUENTES DE DATOS EN ESPAÑA (INGRESO Y GASTO)**

- **Encuestas de Presupuestos Familiares (EPF)**
  - **Panel de Hogares de la Unión Europea (PHOGUE)**
  - **Encuesta de Condiciones de Vida (ECV)**
  - **Encuesta Financiera de las Familias (EFF)**
  - **Paneles de IRPF del IEF**
  - **Muestras IEF-AEAT**
  - **Muestra continua de vidas laborales (SS)**
- 



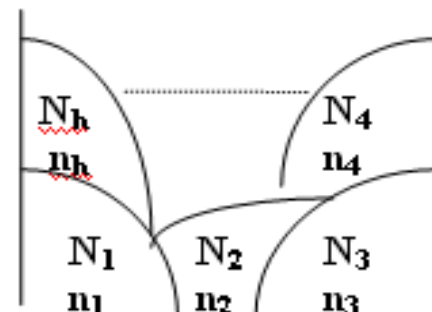
# MUESTRAS IRPF

<b>Población objetivo</b>	<b>Declaraciones presentadas en el IRPF</b>
<b>Ámbito geográfico</b>	<b>Territorio de Régimen Fiscal Común</b>
<b>Ámbito temporal</b>	<b>Ejercicios 2002 a 2012</b>
<b>Unidad de muestreo</b>	<b>Declaraciones de IRPF</b>
<b>Marco</b>	<b>Declaraciones cuyo documento de ingreso o devolución es el de IRPF Modelo 100</b>
<b>Tipo de muestreo</b>	<b>Estratificado aleatorio con tres niveles de estratificación: 49 provincias, 12 tramos de renta y dos tipos de declaración (Individual y Conjunta), lo que hace un total de 1.176 estratos</b>
<b>Tamaño muestral</b>	<b>Cerca de 2.000.000 declaraciones</b>
<b>Afijación</b>	<b>El reparto de la muestra en los estratos se ha realizado con afijación de mínima varianza.</b>
<b>Error muestral</b>	<b>&lt;1,5% con un nivel de confianza del 3‰</b>
<b>Contenido</b>	<b>400 variables personales, familiares, fiscales</b>

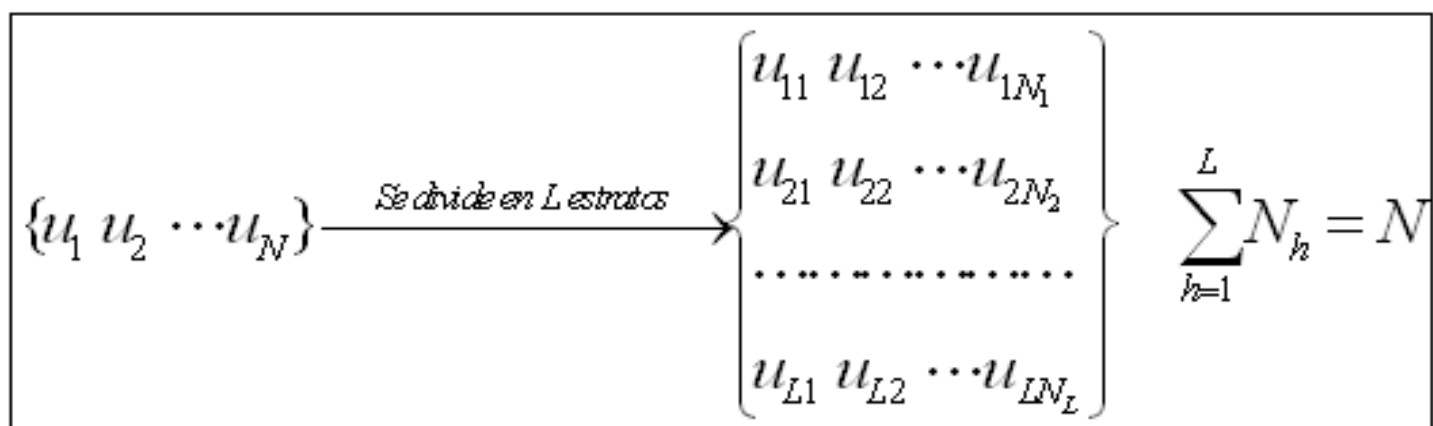
# MUESTREO ESTRATIFICADO

- Una población heterogénea con  $N$  unidades  $\{u_i\}$   $i=1,2,\dots,N$  se subdivide en  $L$  subpoblaciones lo más homogéneas posibles dentro de sí y heterogéneas entre sí no solapadas denominadas estratos de tamaños  $N_1, N_2, \dots, N_L$ .
- La muestra estratificada de tamaño  $n$  se obtiene seleccionando  $n_h$  elementos ( $h=1,2,\dots,L$ ) de cada uno de los  $L$  estratos en que se subdivide la población de forma independiente por m. aleatorio simple.

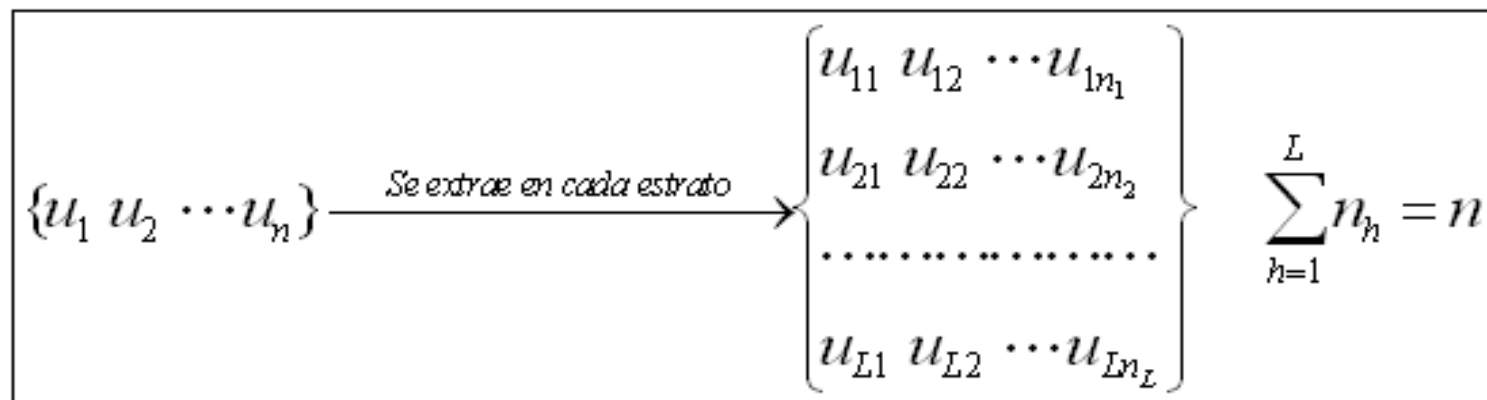




## POBLACIÓN



## MUESTRA



**$49 \times 12 \times 2 = 1176$  estratos**

**P2**

**P1**

**T2**

**P49**

**T12**

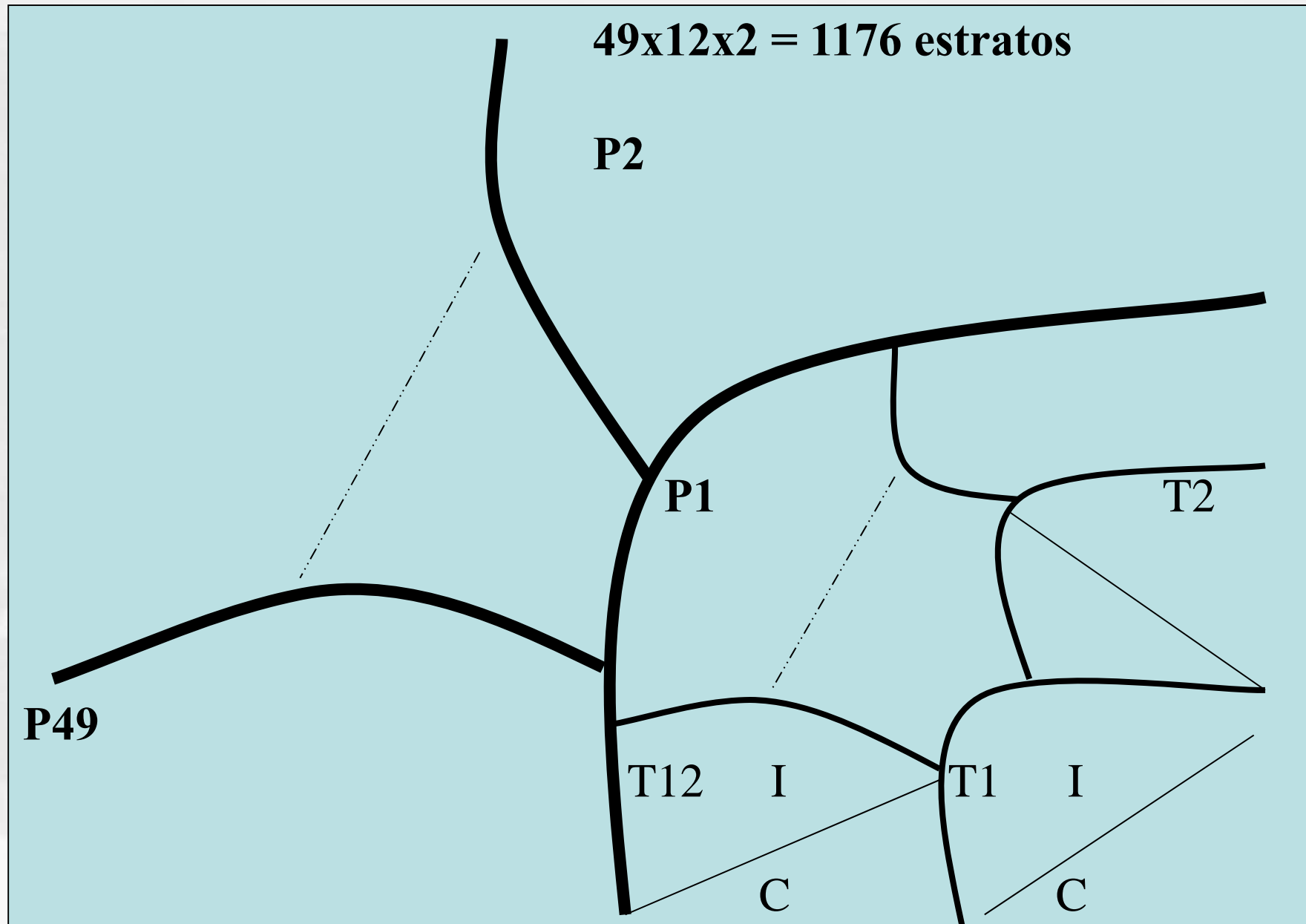
**I**

**T1**

**I**

**C**

**C**



- **Variable utilizada para definir los tramos de renta:**

*Renta = Saldo Neto de Rendimiento e Imputaciones de Renta (var 30) + Parte especial de la B.I. previa a la aplicación del exceso de mínimo exento (var44-var45-var46).*

- **Tramos de renta:**

- Negativas y 0
- Positivas y hasta 6.000 euros
- De 6.001 a 12.000 euros
- De 12.001 a 18.000 euros
- De 18.001 a 24.000 euros
- De 24.001 a 30.000 euros
- De 30.001 a 36.000 euros
- De 36.001 a 42.000 euros
- De 42.001 a 48.000 euros
- De 48.001 a 54.000 euros
- De 54.001 a 60.000 euros
- De más de 60.000 euros.

# RAZONES PARA ESTRATIFICAR

-El muestreo estratificado puede aportar información más precisa de algunas subpoblaciones que varían bastante en tamaño y propiedades entre sí, pero que son homogéneas dentro de sí

-El uso adecuado del muestreo estratificado puede generar ganancia en precisión. El error total derivado del muestreo en todos los estratos se observa que es menor que en el caso de no estratificar la población


- **Conveniencias de tipo administrativo**

Por ejemplo, en el caso de agencias u organismos públicos que disponen de sucursales en distintos puntos, cada una de las cuales supervisaría la encuesta en su correspondiente estrato poblacional, con el consiguiente ahorro en costes de organización, desplazamientos, etc.

- **Requerimiento de estimaciones para ciertas áreas o regiones geográficas**

# ESTIMADORES

- El estimador de cualquier total poblacional en muestreo estratificado aleatorio es la suma de los estimadores del total en cada estrato. Se tiene:

$$\hat{X}_{st} = \sum_{h=1}^L \hat{X}_h = \sum_{h=1}^L N_h \bar{x}_h = \sum_{h=1}^L \frac{N_h}{n_h} x_h = \sum_{h=1}^L fe_h x_h$$


- Por lo tanto, para estimar cualquier total poblacional se suman productos de los FACTORES DE ELEVACIÓN por los totales muestrales en cada estrato (se elevan tamaños y variables)

- El estimador de cualquier media es la media ponderada de los estimadores de la media en cada estrato, siendo los coeficientes de ponderación  $W_h = N_h/N$  ( $N_h$  es el tamaño poblacional del estrato y  $N$  es el tamaño de la población = 15481382 declaraciones).

$$\hat{\bar{X}}_{st} = \bar{x}_{st} = \sum_{h=1}^L W_h \bar{x}_h = \sum_{h=1}^L \underbrace{\frac{N_h}{N}}_{W_h} \frac{1}{n_h} x_h = \frac{1}{N} \sum_{h=1}^L \frac{N_h}{n_h} x_h = \frac{1}{N} \sum_{h=1}^L f e_h x_h$$

- Por lo tanto, para estimar cualquier media poblacional se suman los productos de los FACTORES DE ELEVACIÓN por los totales muestrales en cada estrato y se divide por el tamaño poblacional.

# ERRORES DE ESTIMACIÓN

## Errores absolutos

$$\hat{V}(\hat{X}_{st}) = \sum_{h=1}^L N_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}, \quad \hat{V}(\bar{X}_{st}) = \sum_{h=1}^L W_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}$$

$\hat{S}_h^2$  = *cuasivarianza muestral en el estrato h*

$$f_h = n_h / N_h = 1/f_{eh}$$

## Errores relativos

$$\hat{C}_v(\hat{X}_{st}) = \frac{\sqrt{\hat{V}(\hat{X}_{st})}}{\hat{X}_{st}}$$

$$\hat{C}_v(\bar{x}_{st}) = \frac{\sqrt{\hat{V}(\bar{x}_{st})}}{\bar{x}_{st}}$$

# **AFIJACIÓN DE LA MUESTRA**

**El reparto de la muestra en los estratos se ha realizado mediante afijación de mínima varianza**

**La afijación de mínima varianza o afijación de Neyman consiste en determinar los valores de  $n_h$  (número de unidades que se extraen del estrato  $h$ -ésimo para la muestra) de forma que para un tamaño de muestra fijo igual a  $n$  la varianza de los estimadores sea mínima.**

$$\left. \begin{array}{l} \min V(\bar{x}_{st}) \\ \sum_{h=1}^L n_h = n \end{array} \right\}$$

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^L N_h S_h} = n \cdot \frac{\frac{N_h}{N} S_h}{\sum_{h=1}^L \frac{N_h}{N} S_h} = n \cdot \frac{W_h S_h}{\sum_{h=1}^L W_h S_h}$$

# ¿PORQUÉ ESTA AFIJACIÓN?

- La utilidad de la afijación de mínima varianza es mayor si hay grandes diferencias en la variabilidad en los estratos, que es el caso que nos ocupa.
- En otro caso la mayor sencillez y autoponderación de la afijación proporcional hacen preferible el empleo de ésta.

- Vemos que los valores de  $n_h$  son proporcionales a los productos  $N_h \cdot S_h$  y en el supuesto de que  $S_h = S, \forall h$  esta afijación de mínima varianza coincidiría con la proporcional, tal y como se ve a continuación:

$$S_h = S \Rightarrow n_h = n \cdot \frac{N_h S}{\sum_{h=1}^L N_h S} = \frac{n N_h}{N} = k N_h \text{ con } k = \frac{n}{N}$$

- **El muestreo estratificado con afijación proporcional es más preciso que el muestreo aleatorio simple**, produciéndose la igualdad de precisiones cuando las medias de los estratos son todas iguales.
- **El muestreo estratificado con afijación de mínima varianza es más preciso que el muestreo estratificado con afijación proporcional**, produciéndose la igualdad de precisiones cuando las cuasidesviaciones típicas de los estratos son todas iguales.

$$V_{MAS}(\bar{x}) \geq V_{MEP}(\bar{x}) \geq V_{MEMV}(\bar{x})$$

- El muestreo estratificado con afijación de mínima varianza es más preciso que el muestreo estratificado con afijación proporcional y que el aleatorio simple, siendo además el estratificado con afijación proporcional más preciso que el aleatorio simple.

# TAMAÑO DE LA MUESTRA

Error realtivo de muestreo < 1,5% y  
coeficiente de confianza < 3 por mil

$$n = \frac{\lambda_{\alpha}^2 \left( \sum_{h=1}^L N_h S_h \right)^2}{e_{r\alpha}^2 N^2 \bar{X}^2 + \lambda_{\alpha}^2 \sum_{h=1}^L N_h S_h^2} \cong 907399$$

**Tamaño entre 900.000 y 950.000**

# TAMAÑO DE LA MUESTRA

- El tamaño muestral se ha obtenido para un error de muestreo menor del 1,5% con un nivel de confianza del 3 por mil.
- Se han seleccionado para la muestra 907.399 declaraciones con información relativa a 225 variables. Las 200 primeras variables (var1-var200) se corresponden con las casillas del mismo número en el impreso de declaración. Las restantes variables contienen la siguiente información:

# VARIABLES

<b>Factor</b>	Factor de elevación de la muestra
<b>EstCv</b>	Estado civil de declarante.
<b>Sexo</b>	Sexo del declarante
<b>LiModelo</b>	Modelo de declaración
<b>DEC</b>	Tipo de declaración
<b>EjnacD</b>	Ejercicio de nacimiento del declarante
<b>EjnacC</b>	Ejercicio de nacimiento del conyuge
<b>MinusD</b>	Grado de minusvalía del declarante
<b>MinusC</b>	Grado de minusvalía del conyuge
<b>NmDesc</b>	Número total de descendientes
<b>NmDesc0</b>	Número de descendientes <3 años
<b>NmDesc3</b>	Número de descendientes $\geq 3$ Y $< 16$ años
<b>NmDesc1618</b>	Número de descend $\geq 16$ Y $< 18$ años
<b>NmDesc1825</b>	Número de descend $\geq 18$ Y $< 25$ años
<b>NmDescR</b>	Número de descendientes $\geq 25$ años

# VARIABLES

<b>NmDescD</b>	Número de descendientes con edad desconocida
<b>NmDesM0</b>	Número de descendientes sin minusvalía.
<b>NmDesM33</b>	Número de descend con minusvalía $> 0$ Y $< 33$ %
<b>NmDesM65</b>	Número de desc con minusvalía $\geq 33$ Y $< 65$ %
<b>NmDesMR</b>	Número de descendientes con minusvalía $\geq 65$ %
<b>NmDiscD</b>	Número de descendientes con minusvalía
<b>NmAsc</b>	Número de ascendientes
<b>NmDiscA</b>	Número de ascendientes con minusvalía
<b>NmM65A</b>	Número de ascendientes $> 65$ años
<b>RENTA</b>	Variable de renta utilizada para la estratificación. Es igual a $\text{var30} + \text{var44} - \text{var45} - \text{var46}$
<b>CCAA</b>	Comunidad Autónoma

# VARIABLES

<b>•</b>	<b><u>Var</u></b>	<b><u>Casilla</u></b>	<b><u>Concepto</u></b>
•	1	0	Rendimientos del trabajo.Retribuciones dinerarias. Ingresos íntegros
•	210	1	Rendimientos del trabajo.Retribuciones en especie.Valoración
•	211	2	Rendimientos del trabajo.Retribuciones en especie.Ingresos a cuenta
•	212	3	Rendimientos del trabajo.Retribuciones en especie.Ingresos a cuenta repercutidos
•	2	4	Rendimientos del trabajo.Retribuciones en especie (excepto contribuciones empresariales a Planes de Pensiones y Mutualidades de Previsión Social)
•	3	5	Rendimientos del trabajo.Contribuciones empresariales a Planes de Pensiones y Mutualidades de Previsión Social: importes que se computan al contribuyente
•	4	6	Rendimientos del trabajo.Reducciones especiales (art.17.2 Ley 40/1998)
•	213	7	Rendimientos del trabajo.Cotizaciones a la Seguridad Social o a Mutualidades Generales de Funcionarios, detracciones por derechos pasivos y cotizaciones a Colegios de Huérfanos o entidades similares
•	214	8	Rendimientos del trabajo.Cuotas satisfechas a sindicatos
•	215	9	Rendimientos del trabajo.Cuotas satisfechas a colegios profesionales (si la colegiación es obligatoria y con un máximo de 300,51 euros anuales)
•	216	10	Rendimientos del trabajo.Gastos de defensa jurídica derivados directamente de litigios con el empleador (máximo: 300,51 euros anuales)
•	5	11	Rendimientos del trabajo.Total gastos deducibles
•	220	12	Rendimientos del trabajo.Rendimiento neto

# REPRESENTATIVIDAD (CALIDAD)

Concepto	Badespe	Muestra 2002	Diferencia
Rendimientos íntegros del trabajo	243.075.66	242.789.154	-0,12%
Rendimientos netos del trabajo	232.639.55	231.832.106	-0,35%
Rendimientos íntegros del capital inmobiliario	9.125.805	9.086.934	-0,43%
Rendimientos netos del capital inmobiliario	7.015.814	6.997.485	-0,26%
Rendimientos íntegros del capital mobiliario	11.946.944	11.988.305	0,35%
Rendtos netos reducidos del capital mobiliario	11.438.727	11.441.949	0,03%
Rendimientos netos de actividades económicas en estimación directa normal	4.453.808	4.302.413	-3,40%
Rendtos netos de actividades económicas en estimación directa simplificada	12.728.378	12.596.736	-1,03%
Rendimientos netos de actividades económicas en estimación objetiva (excepto agrícolas, ganaderas y forestales)	7.973.453	7.964.426	-0,11%
Rendimientos netos de actividades agrícolas, ganaderas y forestales en estimación objetiva	3.803.629	3.866.743	1,66%
Saldo neto de ganancias y pérdidas patrimoniales $\leq 1$ año	923.890	910.870	-1,41%
Saldo neto de ganancias y pérdidas patrimoniales $> 1$ año	7.841.562	8.183.497	4,36%

# REPRESENTATIVIDAD (CALIDAD)

Concepto	Badespe	Muestra (02)	Diferencia
<u>Mínimos y bases</u>			
Mínimos personales y familiares	74.390.378	74.709.399	0,43%
Parte general de la base imponible	174.495.71	174.160.284	-0,19%
Parte especial de la base imponible	7.841.562	8.098.866	3,28%
Base liquidable general	169.241.05	169.018.760	-0,13%
Base liquidable especial	7.577.976	7.840.513	3,46%
Cuota íntegra estatal	31.766.880	31.797.648	0,10%
Cuota íntegra autonómica	15.646.293	15.661.447	0,10%
<u>Cuotas</u>			
Cuota líquida estatal	28.746.840	28.802.341	0,19%
Cuota líquida autonómica	14.109.815	14.136.455	0,19%
Cuota diferencial	-3.149.754	-3.156.228	0,21%

# LIMITACIONES

Representa sólo a los declarantes del impuesto en el ejercicio. Se complementa con la muestra de no declarantes del mismo año en construcción

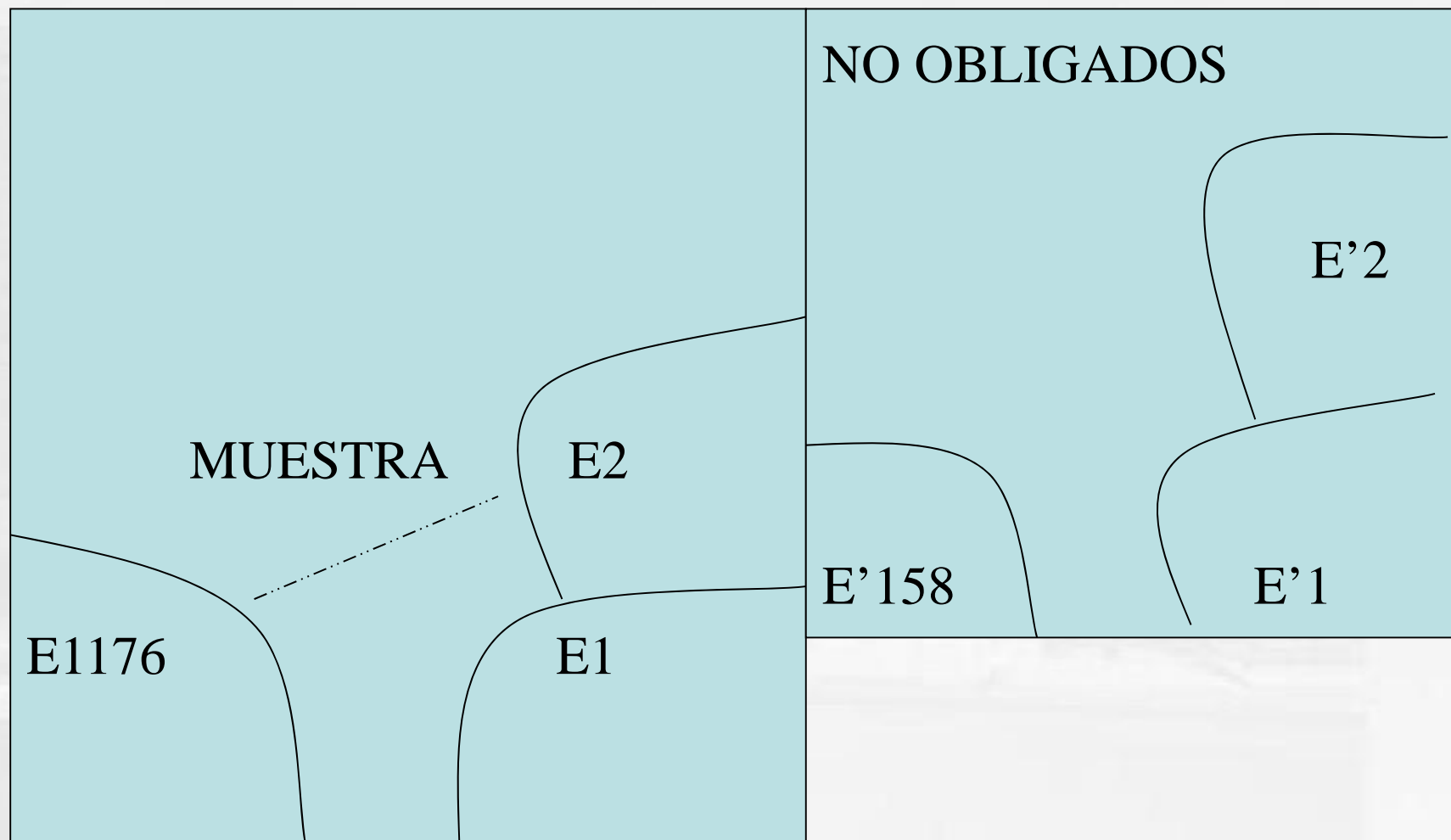
La unidad de análisis es la declaración de IRPF: no identifica a los cónyuges que hayan declarado individualmente, ni permite separar las rentas de los que lo han hecho conjuntamente. No es posible relacionar declaraciones individuales de personas que pertenecen a la misma unidad familiar, ni contribuyentes que viven en el mismo hogar, lo que dificulta los análisis de desigualdad.  
Mejora prevista para el Panel

No proporciona algunos valores monetarios clave del impuesto: no grabados en el ejercicio. El caso más significativo es el de los mínimos personales y familiares.  
La muestra proporciona datos personales y familiares suficientes para simular dichos mínimos.

## **MUESTRAS DE NO OBLIGADOS NO DECLARANTES**

- **Una segunda parte del trabajo tiene como finalidad considerar ahora la muestra de declarantes de IRPF, ampliada con los “no obligados no declarantes” (Modelo 190).**
- **De esta forma se puede contrastar el efecto de este segundo grupo, que habitualmente no se incorpora en los análisis y cuya importancia es esencial dado que se trata de un colectivo que supone el 20% de la población total de contribuyentes del IRPF.**

- La información de los no obligados no declarantes proviene de una muestra de este colectivo con una representatividad similar a la de la muestra de declarantes.
- La información se estratifica también por provincias y por tramos de renta, pero ahora no existe el tercer nivel de estratificación
- Metodológicamente, dado que los estratos no son coincidentes en las dos muestras, uniremos los estratos de la muestra de declarantes con los de la de no obligados no declarantes, obteniendo una muestra final con un número de estratos igual a la suma de los de las dos muestras ( $L+L'$ ).



# POBLACIÓN Y MUESTRA

- Se consideran todas las retenciones que los perceptores han hecho al mismo individuo.
- Si su renta total está entre 8000 y 22000 es seleccionable para la muestra (cumple las condiciones de no obligado)
- Los tramos de renta son [8000, 12000], [12000 18000] [18000, 22000]
- 48 provincias x 3 tramos = 144 estratos
- Para cada individuo seleccionado para la muestra se consideran todos sus 190

# VARIABLES 190

- **[var1]** Rendimientos del trabajo.Retríbuciones dinerarias. Ingresos íntegros
- **[var2]** Rendimientos del trabajo.Retríbuciones en especie (excepto contribuciones empresariales a Planes de Pensiones y Mutualidades de Previsión Social)
- **[var4]** Rendimientos del trabajo.Reducciones especiales (art.17.2 Ley 40/1998)
- **[var40]** Reducción por pensiones compensatorias al cónyuge y anualidades por alimentos (excepto en favor de los hijos) por decisión judicial.Importe de la reducción
- **[var50]** Importe de las anualidades por alimentos en favor de los hijos satisfechas por resolución judicial
- **[var101]** Importe de la retención (debe ser similar a la cuota resultante de la autoliquidación)

# VARIABLES 190

- **[DEC]** Tipo Declaración
- **[NmDesc0]** Número de hijos menores de 3 años
- **[NmDesc3]** Número de hijos entre 3 y 16 años
- **[NmDescR]** Número de hijos mayores de 25 años
- **[NmDescD]** Número de hijos con edad desconocida
- **[NmDesM0]** Número de hijos sin minusvalía
- **[NmDesM65]** Número de hijos con minusvalía superior al 33% e inferior al 65%
- **[NmDesMR]** Número de hijos con minusvalía superior al 65%
- **[NmDiscD]** Número de hijos con grado de minusvalía desconocida
- **[EstCv]** Estado Civil
- **[NmDesc]** Número de hijos totales
- **[DiscapTot]** Número de hijos discapacitados
- **[NmDescMenor]** Número de hijos menores de 18 años

# VARIABLES 190

- **[Factor]** Factor de elevación
- **[prov]** Provincia
- **[CA]** Comunidad Autónoma
- **[NmAsc]** Número de ascendientes
- **[NmDiscA]** Número de ascendientes con minusvalía
- **[NmM65A]** Número de ascendientes > 65 años
- **[NmAscDeduc]** Número de ascendientes con derecho a deducción
- **[Minus33]** Trabajador con Minusvalía superior al 33%
- **[Minus33Plus]** Trabajador con Minusvalía superior al 33% y movilidad reducida
- **[Minus65]** Trabajador con Minusvalía superior al 65%

# ESTIMADORES

- Factores de elevación de las dos muestras se mantienen en la muestra final
- Las variables son coincidentes en las dos muestras
- El estimador de cualquier total poblacional  $X$  en muestreo estratificado aleatorio es ahora la suma de los estimadores del total en cada uno de los  $L+L'$  estratos. Se tiene:

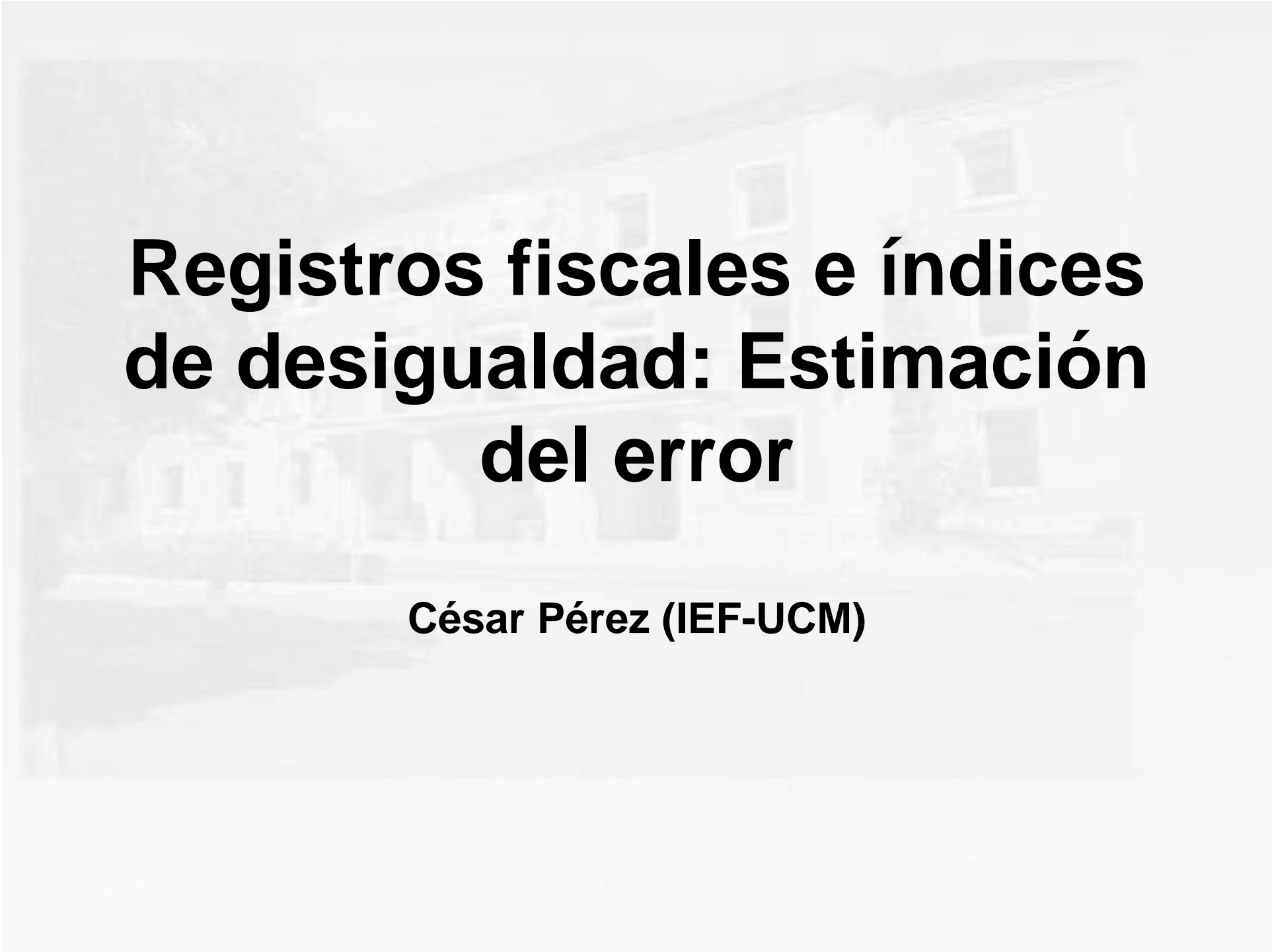
$$\hat{X}_{st} = \sum_{h=1}^{L+L'} \hat{X}_h = \sum_{h=1}^{L+L'} N_h \bar{x}_h = \sum_{h=1}^{L+L'} \frac{N_h}{n_h} x_h = \sum_{h=1}^{L+L'} fe_h x_h$$

$$\left\{ \begin{array}{l} \bar{x}_h = \text{media muestral en el estrato } h \\ x_h = \text{total muestral en el estrato } h \\ N_h = \text{tamaño poblacional del estrato } h \\ n_h = \text{tamaño muestral del estrato } h \\ fe_h = \text{factor de elevación del estrato } h \end{array} \right.$$

**Igual que en la muestra de declarantes, para estimar cualquier total poblacional se suman los productos de los factores de elevación  $f_{eh}$  por los totales muestrales en cada estrato  $x_h$ .**

**El estimador de cualquier media en muestreo estratificado aleatorio es ahora la media ponderada de los estimadores de la media en cada estrato, siendo los coeficientes de ponderación  $W_h = N_h/(N+N')$  de suma unitaria ( $N_h$  es el tamaño poblacional del estrato,  $N$  es el tamaño de la población de declaraciones y  $N'$  es el tamaño de la población de no obligados no declarantes)**

$$\hat{X}_{st} = \bar{x}_{st} = \sum_{h=1}^{L+L'} W_h \bar{x}_h = \sum_{h=1}^{L+L'} \underbrace{\frac{N_h}{N+N'}}_{W_h} \frac{1}{n_h} x_h = \frac{1}{N+N'} \sum_{h=1}^{L+L'} \frac{N_h}{n_h} x_h = \frac{1}{N+N'} \sum_{h=1}^{L+L'} f e_h x_h$$

A faded, grayscale background image of a multi-story building with a tiled roof and several windows, serving as a backdrop for the title text.

# **Registros fiscales e índices de desigualdad: Estimación del error**

**César Pérez (IEF-UCM)**

**Disponibilidad de muestras transversales de declarantes del IRPF, elaboradas por el Instituto de Estudios Fiscales y la Agencia Tributaria.**

**Permiten analizar diversos aspectos relacionados con la desigualdad de la renta y el papel redistributivo del IRPF.**

**Las muestras de declarantes de IRPF son un instrumento adecuado para realizar análisis formal de desigualdad y redistribución.**

**Permiten realizar adecuadamente las tareas de microsimulación para analizar medidas de política fiscal**

**Lo habitual es calcular los índices de Gini (IG) antes y después de la aplicación del impuesto**

**Calcular el índice de Reynolds-Smolensky (IRS), que expresa el grado de redistribución del impuesto así como la diferencia de los dos IG mencionados**

**El índice de Kakwani (1977) (IK), que mide la progresividad del impuesto mediante la diferencia entre el IG de la renta antes de impuestos y un índice de concentración de las cuotas líquidas ordenadas según renta. .**

**Todo está en función del índice de Gini**

**El índice de Gini es una medida de concentración relativa definida como la mitad de la diferencia media para cada par de observaciones de renta, dividida por el valor media de la variable cuya distribución se evalúa, tradicionalmente expresado como:**

$$G(y) = \frac{\sum_{i=1}^n \sum_{j=1}^n |y_i - y_j|}{n^2 \bar{y}}$$

**Siguiendo a Glasser (1962) y Dixon (1987),  
alternativamente la fórmula del coeficiente de Gini  
puede escribirse como sigue:**

$$G(y) = \frac{1}{n(n-1)\bar{y}} \sum_{i=1}^n (2i - n - 1) y_i$$

**Adicionalmente, cuando se dispone de una muestra de tamaño n extraída de una población de tamaño N, el índice de Gini poblacional puede estimarse insesgadamente mediante el estimador siguiente:**

$$\hat{G}(y) = \frac{1}{N\bar{y}} \sum_{i=1}^n K_i y_i \left( 2 \sum_{j=1}^n K_j - K_i - N \right)$$

**donde y es la variable renta, n es el tamaño de la muestra,  $K_i$  es el factor de elevación y N es el tamaño poblacional**

# CUANTIFICACIÓN DEL ERROR AL ESTIMAR EL ÍNDICE DE GINI

Habitualmente el error absoluto de un estimador insesgado suele medirse, a partir de los datos de una muestra, mediante la estimación de su varianza.

Pero el problema aparece al intentar estimar la varianza cuando la expresión del estimador es complicada, tal y como ocurre en el caso del estimador del Índice de Gini.

En estas situaciones se acude a los métodos específicos de estimación de varianzas utilizados en la teoría del muestreo.

Entre estos métodos tenemos el método de las muestras interpenetrantes, el método de los grupos aleatorios, el método de las semimuestras reiteradas, el método de Jackknife y el método Bootstrap

# INTERVALOS DE CONFIANZA PARA EL ÍNDICE DE GINI

Normalidad

$$\left[ \hat{\theta} - \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}}, \hat{\theta} + \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}} \right]$$

No normalidad

$$\left[ \hat{\theta} - \lambda_{\alpha} \sigma(\hat{\theta}), \hat{\theta} + \lambda_{\alpha} \sigma(\hat{\theta}) \right]$$

# MUESTRAS INTERPENETRANTES

Se utiliza cuando tenemos un conjunto de dos o más muestras, elegidas con el mismo esquema de muestreo (independientes o no) y tales que cada una proporcione una estimación válida del parámetro que se pretenda estimar con el mismo error de muestreo.

Sean  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$  estimadores insesgados de  $\theta$  basados en  $k$  muestras independientes. Su media:

$$\hat{\theta} = \frac{1}{k} \sum_i^k \hat{\theta}_i$$

es también un estimador insesgado de  $\theta$

# SUBMUESTRAS INTERPENETRANTES

Se utiliza cuando tenemos un conjunto de dos o más muestras, elegidas con el mismo esquema de muestreo (independientes o no) y tales que cada una proporcione una estimación válida del parámetro que se pretenda estimar con el mismo error de muestreo.

Sean  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$  estimadores insesgados de  $\theta$  basados en  $k$  muestras independientes. Su media:

$$\hat{\theta} = \frac{1}{k} \sum_i^k \hat{\theta}_i$$

es también un estimador insesgado de  $\theta$

Un estimador insesgado para la varianza es :

$$\hat{V}(\hat{\theta}) = \frac{1}{k(k-1)} \left( \sum_i^k \hat{\theta}_i^2 - k\hat{\theta}^2 \right)$$

El estimador de la varianza para el índice de Gini será:

$$\hat{V}(\hat{G}) = \frac{1}{k(k-1)} \left( \sum_i^k \hat{G}_i^2 - k\hat{G}^2 \right)$$

$$\hat{G} = \frac{1}{k} \sum_i^k \hat{G}_i$$

En nuestro caso usamos 30 muestras independientes de tamaño 30.000 declaraciones del IRPF. El estimador de la varianza para el índice de Gini es:

$$\hat{V}(\hat{G}) = \frac{1}{k(k-1)} \left( \sum_i^k \hat{G}_i^2 - k\hat{G}^2 \right) = 0,00000293063$$

$$\hat{C}_v(\hat{G}) = \frac{\sqrt{\hat{V}(\hat{G})}}{\hat{G}} = 0,001295489$$

El error relativo para el estimador del índice de Gini es del 0,1295%, (uno por mil). Resultado óptimo derivado del elevado tamaño de las muestras, del elevado número de muestras y de su elevada precisión

# GRUPOS ALEATORIOS

Se extrae una muestra de  $n$  unidades de una población de tamaño  $N$ . Dicha muestra se subdivide en  $K$  submuestras de igual tamaño  $m$ , de modo que  $n=K.m$ . Estas submuestras se denominan grupos aleatorios, y además de ser submuestras de la muestra, también son muestras de la población completa.

En estas condiciones si  $\hat{\theta}$  es un estimador insesgado de la característica poblacional  $\theta$  basado en la muestra completa  $W$ , y si  $\hat{\theta}_r$  es un estimador insesgado de la característica poblacional  $\theta$  basado en el  $r$ -ésimo grupo aleatorio, un estimador insesgado de la varianza de es el siguiente:

$$\hat{V}(\hat{\theta}) = \frac{1}{K(K-1)} \sum_{r=1}^K (\hat{\theta}_r - \hat{\theta})^2$$

**En nuestro caso utilizamos 30 submuestras independientes de tamaño 30.000 declaraciones del IRPF. El estimador de la varianza para el índice de Gini será:**

$$\hat{V}(\hat{G}) = \frac{1}{K(K-1)} \sum_{r=1}^K (\hat{G}_r - \hat{G})^2 = 0,0000029792$$

$$\hat{C}_v(\hat{G}) = \frac{\sqrt{\hat{V}(\hat{G})}}{\hat{G}} = 0,001307335$$

# MÉTODO BOOTSTRAP (AUTOGENRACIÓN)

Para llevarlo a cabo partimos de la muestra de tamaño un millón de declaraciones de IRPF extraída de una población de 16 millones de declaraciones. A continuación extraemos de la muestra inicial  $M=1000$  muestras con reposición, también de tamaño un  $\hat{\theta}_j^*$  millón y calculamos en cada una de ellas el estimador para el cual estamos calculando el error (índice de Gini). La precisión del estimador se obtiene por la expresión:

$$\hat{\sigma}_{BOOT\ 1000} = \sqrt{\frac{\sum_{j=1}^M (\hat{\theta}_j^*)^2 - \left(\sum_{j=1}^M (\hat{\theta}_j^*)\right)^2 / M}{M - 1}} = 0,00052414$$

$$\hat{C}_V(\hat{G}) = \frac{\hat{\sigma}_{BOOT1000}}{\hat{G}} = 0,00125543$$

**Si ahora consideramos M=5000 muestras, tenemos los siguientes resultados:**

$$\hat{\sigma}_{BOOT5000} = \sqrt{\frac{\sum_{j=1}^M (\hat{\theta}_j^*)^2 - \left( \sum_{j=1}^M (\hat{\theta}_j^*) \right)^2 / M}{M-1}} = 0,00052414$$

$$\hat{C}_V(\hat{G}) = \frac{\hat{\sigma}_{BOOT5000}}{\hat{G}} = 0,00122569$$

# NORMALIDAD

Método de las submuestras interpenetrantes.

El intervalo del índice de Gini es  $0,4175 \pm 0,001061$

Método de los grupos aleatorios: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,001069$

Método Bootstrap con 1000 muestras: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,001027$

Método Bootstrap con 5000 muestras: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,001003$

# NO NORMALIDAD

Método de las submuestras interpenetrantes.

El intervalo del índice de Gini es  $0,4175 \pm 0,00242$

Método de los grupos aleatorios: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,00244$

Método Bootstrap con 1000 muestras: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,00234$

Método Bootstrap con 5000 muestras: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,00228$

## ALGUNAS CONCLUSIONES

Los errores obtenidos para las estimaciones del índice de Gini con la muestra son pequeños.

Este hecho indica la precisión de los cálculos de los índices de desigualdad, progresividad y redistribución basados en la muestra.

De aquí se deriva la **pertinencia del uso de estos índices con fiabilidad en las simulaciones del impuesto basadas en las muestras del IRPF**

Por otro lado, se observa que los intervalos son más anchos cuando no hay normalidad (menos precisión en la estimación del índice de Gini).

Asímismo, se observa que el método menos afectado por la falta de normalidad en cuanto a precisión según los intervalos de confianza es el método Bootstrap.

Esto corrobora la idea de que el bootstrapping es una técnica muy adecuada como método especial para la estimación de varianzas en estimadores complejos.

# **LA MUESTRA DE IRPF DE 2009: DESCRIPCIÓN GENERAL Y PRINCIPALES MAGNITUDES**

*Autores: César Pérez López*

*María Jesús Burgos Prieto*

*Sara Huete*

*Carmen Gallego*

**Instituto de Estudios Fiscales**

**DOC. n.º 11/2012**

# **PANEL DE DECLARANTES DE IRPF DEL IEF 1982-1998**

- **Diseño**
  - **Soporte informático**
  - **Limitaciones**
  - **Novedades y proyectos**
- 



## Diseño

**Iniciado en 1988: año base 1987;  
planificación inicial para más-menos 5 años  
(1982-1992)**

**Cortes transversales: muestreo aleatorio  
simple por declaraciones (1 de cada 50; por  
provincia; territorio fiscal común)**

**Año base: representatividad del 99,5% (95%  
por CC.AA. Con problemas en las menos  
pobladas: Rioja; Cantabria; Murcia; etc.)**

**Pensado inicialmente para tributación  
conjunta de matrimonios**

**Extensión del Panel Puro + muestreo aleatorio  
de altas**



# DISEÑO MUESTRAL

- **Ámbito: Poblacional, geográfico y temporal**

La población objetivo son las declaraciones presentadas del Impuesto sobre la Renta de las Personas Físicas (IRPF) correspondientes a cada ejercicio del panel. El ámbito geográfico lo constituye el Territorio de Régimen Fiscal Común. El ámbito temporal es el comprendido desde 1982 a 1998.

- **Unidad de muestreo:** Declaraciones (individuales, conjuntas o separadas)

- **Marco:** El marco lo constituyen el conjunto de unidades de entre las cuales se selecciona efectivamente la muestra. Se ha utilizado el marco de lista de declaraciones individuales, separadas y conjuntas de los impresos ordinario, simplificado y abreviado.

# Diseño muestral

- **Tipo de muestreo**

Se ha utilizado muestreo aleatorio simple del 2% eligiendo una declaración de cada cincuenta en cada Delegación de Hacienda.

- **Variables en el panel**

El contenido del panel incluye para cada año las declaraciones (todas sus casillas) de los declarantes que están presentes en el panel ese año. Las variables (campos) que caracterizan el panel son:

*Id\_Panel*: Identificador de cada hogar.

*TipoDeclarante*: Muestra si el declarante que forma parte del panel es principal (P) o cónyuge (C).

# Diseño muestral

- *Id\_TipoDeclaración*: Identifica el tipo de declaración que se ha utilizado. Los posibles códigos son los siguientes:
  - 0→Individual
  - 1→Separada
  - 2→Conjunta
- *Id\_TipoImpreso*: Identifica el tipo de impreso que se ha utilizado en la declaración. Los posibles códigos son los siguientes:
  - 0→Ordinaria
  - 1→Simplificada
  - 3→Abreviada

# Diseño muestral

- *Comunidad*: muestra el código de la comunidad en la que está encuadrada la delegación.
- *dh*: Identifica el código de la delegación de hacienda asignado al declarante.
- *Municipio*: Identifica el código postal del municipio donde reside el declarante.
- *C1, C2, ....., Cn* : Valor de cada una de las casillas en la declaración. Cada año tiene un número distinto de casillas en la declaración.

1982	C1...C99	123.599
1983	C1...C99	130.500
1984	C1...C99	134.957
1985	C1...C99	145.664
1986	C1...C99	165.303
1987	C1...C99	173.979
1988	C1...C99	193.444
1989	C1...C99	208.808
1990	C1...C100	235.646
1991	C1...C100	251.197
1992	C1...C103	277.733
1993	C1...C100	287.291
1994	C1...C104	313.116
1995	C1...C104	325.039
1996	C1...C110	310.859
1997	C1...C126	308.736
1998	C1...C126	308.558

# ESTIMADORES

- El estimador insesgado de cualquier total poblacional para un año del panel en muestreo aleatorio simple es la expansión de la media muestral mediante el tamaño poblacional.

Tenemos:

$$\hat{X} = N\bar{x} = N \frac{1}{n} \sum_{h=1}^n X_i = \frac{N}{n} \sum_{h=1}^n X_i = \frac{N}{n} \times (Total\ muestral) = f_e \times (Total\ muestral)$$

# ESTIMADORES

- Por lo tanto, para estimar cualquier total poblacional se multiplica el factor de elevación por el correspondiente total muestral, siendo el factor de elevación el cociente entre el tamaño poblacional y el muestral (y cuyo valor es aproximada 50 debido a que la muestra es del 2% cada año).

# ESTIMADORES

**El estimador insesgado de cualquier media poblacional para un año del panel en muestreo aleatorio es la media muestral.**

$$\hat{\bar{X}} = \bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

# ERRORES DE ESTIMACIÓN

- Los errores relativos estimados se calculan mediante las expresiones:

$$\hat{C}_v(\hat{X}) = \frac{\sqrt{\hat{V}(\hat{X})}}{\hat{X}}$$

$$\hat{C}_v(\bar{x}) = \frac{\sqrt{\hat{V}(\bar{x})}}{\bar{x}}$$

$$\hat{V}\left(\frac{\hat{X}}{N}\right) = \frac{1}{N^2} \hat{V}(\hat{X}) = \frac{1}{N^2} N^2 (1-f) \frac{\hat{S}^2}{n} = (1-f) \frac{\hat{S}^2}{n}$$

# DESGASTE (ATTRITION)

- Se trata de un panel “expandido” en contraposición con la idea más habitual de paneles puros. La razón de esta decisión es que estos últimos no constituyen una fiel representación transversal de la población de la que se extraen porque sobre la misma inciden, a lo largo del tiempo, un flujo de entradas y salidas que, en nuestro caso, se materializaría, respectivamente, en los nuevos contribuyentes y en los que dejan de serlo en cada ejercicio.

# DESGASTE

- Estadísticamente ello significa que **dichos paneles puros no constituirían, cada año, una muestra aleatoria de los contribuyentes del año.** Ello tendrá efectos perniciosos en dos vertientes, cuando menos:

Por una parte, los estudios transversales realizados sobre el panel exigirían el **desarrollo de estimadores estadísticos distintos cada año basados en técnicas de reescalamiento muestral**, es decir, de asignación de pesos específicos anuales a cada individuo del panel. Evidentemente, como es natural, dichos estimadores serán tanto más complejos cuanto menor sea el grado de similitud entre el panel puro y el total de declarantes del ejercicio bajo consideración. En consecuencia, las diferencias entre los estimadores de uno u otro año dependerán del grado o nivel de similitud (Panel/Población de contribuyentes) previamente aludido.

# DESGASTE

- Por otra parte, los paneles puros tampoco permitirían analizar las peculiaridades específicas de los nuevos contribuyentes ni de los que dejan de serlo en cada ejercicio, ya que, por su propia naturaleza, no podrían contener ninguno de ellos. En este sentido, esta limitación restaría operatividad al panel, máxime si tenemos en cuenta que en los últimos años los nuevos contribuyentes del IRPF han representado en cada ejercicio aproximadamente un 12% del total y que además, por lógica, deberán existir discrepancias significativas entre los nuevos contribuyentes y los antiguos.

# DESGASTE

- El panel “expandido” surgió, en su momento, como una vía que trataba de eliminar los perniciosos efectos antes aludidos. **Permite sin embargo, llevar a cabo todos los estudios asociados a los paneles puros, ya que contiene un o de ellos como subpanel.**
- Consiste, en esencia, en muestras representativas de contribuyentes que están solapados y que pueden ser extraídos de forma recurrente, una vez seleccionada la primera, mediante la **incorporación (o expansión) de una submuestra adicional de los nuevos contribuyentes de cada año, frente a los anteriores, con una afijación o tamaño al de la muestra ya existente frente al total de declarantes antiguos.**

# DESGASTE

- Obviamente, **los nuevos contribuyentes incorporados al panel continuarán siendo observados desde el momento de su incorporación en adelante.** Las bajas surgirán, de hecho, como resultado de dicha observación, es decir, aparecerán sobre el panel de manera natural cuando no sean encontrados entre los declarantes del año.
- De esta forma, **se dispondrá de muestras representativas de cada ejercicio en las que tendremos identificados los nuevos declarantes y los antiguos así como los que han causado baja en un determinado año, constituyendo la parte común de las mismas el ya mencionado panel puro subcontenido en ellas.**

# ACTUALIZACIÓN

- En este contexto, el panel actualmente existente en el Instituto de Estudios Fiscales deberá ser actualizado año a año como el objetivo de preservar su validez con continuidad. **La finalidad de esta actualización anual es que debemos disponer de un conjunto de contribuyentes observados a lo largo del tiempo que se expanden año a año para identificar transversalmente aquellos que son nuevos de los que son antiguos.** Los tamaños muestrales de unos y otros deben coincidir anualmente en relación con la población que representan para que todos, en conjunto, constituyan, en cada ejercicio, una muestra aleatoria del total de contribuyentes de IRPF del año

# Soporte informático

- Inicialmente el Panel estaba en ficheros formato binario
- Comprendía desde el año 1985 a 1998
- Explotación de datos en Lenguaje C
- Nuevo tratamiento informático:  
Base de datos (SQL Server): estructura relacional (a partir de 1999)
- Actualmente (hasta 2011):  
Programación SAS



# Limitaciones:

- Registro administrativo; cambios legales; cambios población
- Características socioeconómicas: sólo las relevantes fiscalmente
- Legislación fiscal y estructura de la declaración
- Información (casillas) grabadas
- Tributación individual-conjunta
- Obligación de declarar
- Envejecimiento de la muestra (Attrition)



# Nuevo Panel

- Base 2003; retrospectivo hasta 1999; extensión futuro; seguimiento desgaste temporal
- Selección por individuos
- Seguimiento unidades familiares
- Inclusión No Declarantes (retenciones)
- Muestreo estratificado: CC.AA/Renta Bruta/Fuente de Renta
- Mínima varianza/error  $< 1,5\%$ /nivel confianza 3 por mil
- Datos Patrimonio



## PANEL DE IRPF 1999/2007 DEL INSTITUTO DE ESTUDIOS FISCALES: OBJETIVO Y PLANTEAMIENTO GENERAL

- Se trata de disponer de un Panel con información de **rentas fiscales de personas y hogares** de una población representativa de los sujetos pasivos de IRPF en el Territorio de Régimen Fiscal Común a lo largo del tiempo.
- Este Panel debe responder al concepto de **Panel expandido**; es decir que anualmente debe incluirse una representación de las altas que se produzcan controlando también las bajas.

- Dado el objetivo perseguido y la información disponible, se considera que la opción más adecuada es la utilización de los **individuos** como **unidad muestral** al ser esta, y no las declaraciones, una unidad homogénea a lo largo del tiempo.
- Ello conduce a que, **en el caso de las declaraciones conjuntas, se deban individualizar para cada uno de los cónyuges las rentas declaradas conjuntamente y así poder realizar la selección de la muestra;** para ello se utilizará la información de la que la AEAT dispone y que de hecho utiliza en la elaboración de los Borradores que envía a los contribuyentes.

- Para los individuos que resulten seleccionados según los criterios que se establecen, se suministrará la **información sobre las imputaciones individuales realizadas así como toda la información de sus declaraciones presentadas por ellos** (sean obligados o no y realicen declaración conjunta o declaración individual).
- Así mismo, **para poder llevar a cabo análisis referido a hogares, se suministrará el mismo conjunto de información referida a sus cónyuges**, siempre que se disponga de la información que permita su identificación como tales cónyuges.

# **FORMACIÓN DEL MARCO DE LISTA DE INDIVIDUOS**

- El marco de lista de individuos, del que se extraerá la muestra, será el conjunto de declarantes de IRPF (Modelo 100)

# VARIABLES DE ESTRATIFICACIÓN

- *Comunidad autónoma de residencia*
- *12 Tramos de Rentas brutas, aproximadas por los Ingresos íntegros, sin deducir por tanto ni los gastos ni las reducciones [con la excepción de las Rentas de actividades económicas en la que se tomarán los Rendimientos Netos]. En términos de las casillas de la declaración la variable vendría definida por:*  
$$01+02+03+07+12+13+18+20+21+22+23+24+25+26+27+28$$
- *Proporción de ingresos del trabajo sobre el total de rentas (>50% y <50%)*

## Tamaño y afijación

- Se utilizará **afijación de mínima varianza**.
- El **tamaño** vendrá definido por un **error menor del 1,5 por ciento, con un nivel de confianza del 3 por mil**.

## Selección de la muestra en el año base

- En cada CCAA (15 estratos) los individuos se agruparán según el tramo de ingresos íntegros que le corresponda (12 tramos). El resultado serán 180 subestratos . Por último se consideran las dos fuentes de renta (>50% de rentas provenientes del trabajo y <50%). Los estratos serán  $180 \times 2 = 360$

- En cada uno de los subestratos definidos según el tramo de renta, se separaran en dos grupos los individuos según el origen de dichas rentas (**más del 50% de los ingresos son del trabajo y el 50% o menos provienen del trabajo**). El resultado serán  $176 \times 2 = 352$  estratos de último nivel en cada uno de los cuales se realiza la extracción aleatoria simple.
- Una vez seleccionados los individuos que formarán parte de la muestra y cuya información, por tanto, vendrá acompañada del correspondiente factor de elevación, **a efecto de análisis se seleccionarán sus cónyuges, hayan realizado declaración (conjunta o individual) o no la hayan realizado pero se disponga de información**

**15x12x2 = 360 estratos**

**C2**

**C1**

**T2**

**C15**

**T12**

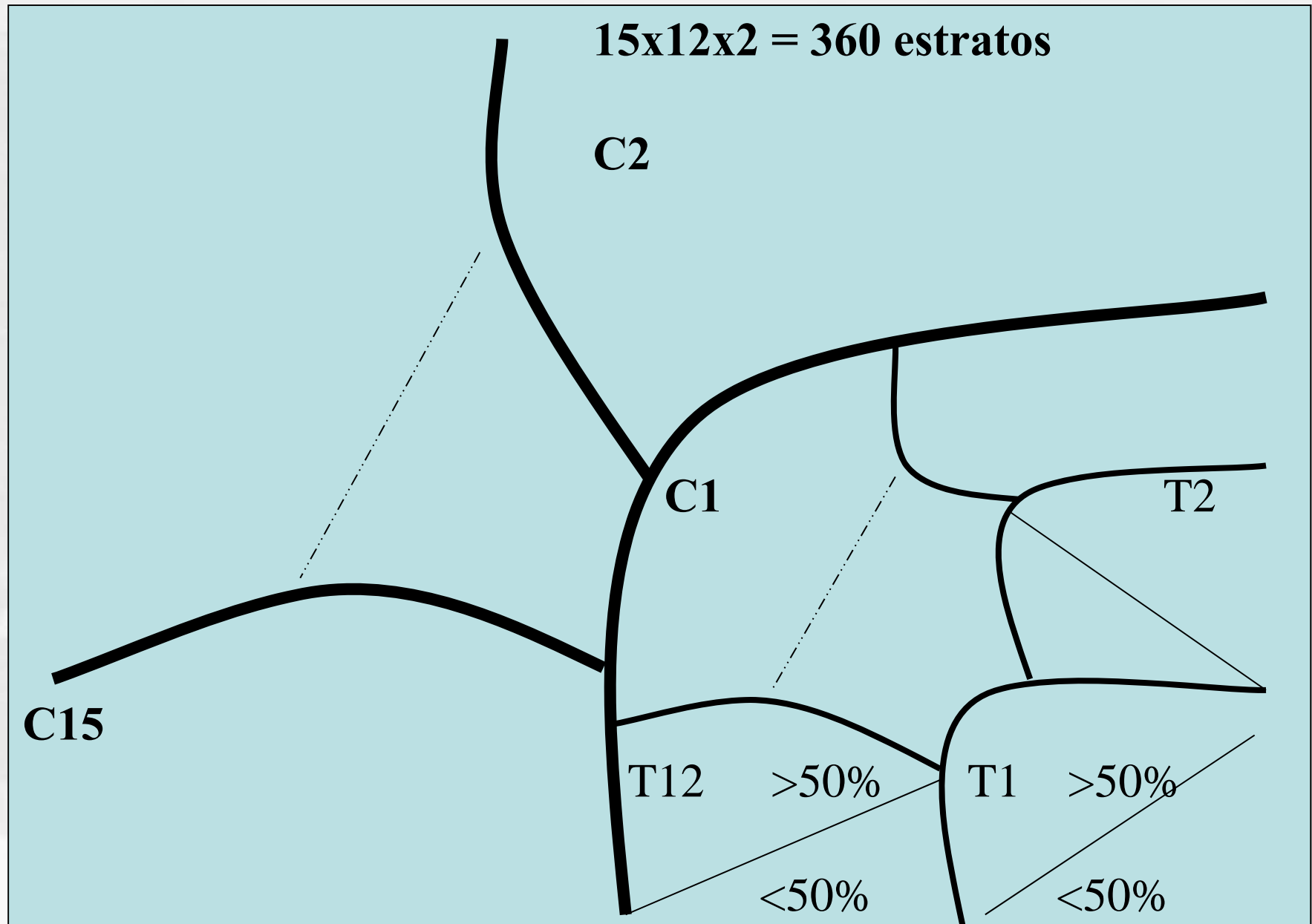
**>50%**

**T1**

**>50%**


**<50%**

**<50%**



# ESTIMADORES

- El estimador de cualquier total poblacional en muestreo estratificado aleatorio es la suma de los estimadores del total en cada estrato. Se tiene:

$$\hat{X}_{st} = \sum_{h=1}^L \hat{X}_h = \sum_{h=1}^L N_h \bar{x}_h = \sum_{h=1}^L \frac{N_h}{n_h} x_h = \sum_{h=1}^L fe_h x_h$$


- Por lo tanto, para estimar cualquier total poblacional se suman productos de los FACTORES DE ELEVACIÓN por los totales muestrales en cada estrato (se elevan tamaños y variables)

- El estimador de cualquier media es la media ponderada de los estimadores de la media en cada estrato, siendo los coeficientes de ponderación  $W_h = N_h/N$  ( $N_h$  es el tamaño poblacional del estrato y  $N$  es el tamaño de la población = 15481382 declaraciones).

$$\hat{\bar{X}}_{st} = \bar{x}_{st} = \sum_{h=1}^L W_h \bar{x}_h = \sum_{h=1}^L \underbrace{\frac{N_h}{N}}_{W_h} \frac{1}{n_h} x_h = \frac{1}{N} \sum_{h=1}^L \frac{N_h}{n_h} x_h = \frac{1}{N} \sum_{h=1}^L f e_h x_h$$

- Por lo tanto, para estimar cualquier media poblacional se suman los productos de los FACTORES DE ELEVACIÓN por los totales muestrales en cada estrato y se divide por el tamaño poblacional.

# TAMAÑO DE LA MUESTRA

Error realtivo de muestreo < 1,5% y  
coeficiente de confianza < 3 por mil

$$n = \frac{\lambda_{\alpha}^2 \left( \sum_{h=1}^L N_h S_h \right)^2}{e_{r\alpha}^2 N^2 \bar{X}^2 + \lambda_{\alpha}^2 \sum_{h=1}^L N_h S_h^2} \cong 400000$$

**Tamaño entre 350.000 y 425.000**

# ERRORES DE ESTIMACIÓN

## Errores absolutos

$$\hat{V}(\hat{X}_{st}) = \sum_{h=1}^L N_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}, \quad \hat{V}(\bar{X}_{st}) = \sum_{h=1}^L W_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}$$

$\hat{S}_h^2$  = *cuasivarianza muestral en el estrato h*

$$f_h = n_h / N_h = 1/f_{eh}$$

## Errores relativos

$$\hat{C}_v(\hat{X}_{st}) = \frac{\sqrt{\hat{V}(\hat{X}_{st})}}{\hat{X}_{st}}$$

$$\hat{C}_v(\bar{x}_{st}) = \frac{\sqrt{\hat{V}(\bar{x}_{st})}}{\bar{x}_{st}}$$

## ***Muestreo en años posteriores al año base (2004 y siguientes)***

- Cada año posterior se debe **extraer una muestra de las altas de individuos** (sea por presentación de declaración sea por información suministrada por los retenedores). Para ello es preciso:
- Detectarlas por comparación con la población de individuos del año anterior.
- Dividir las rentas en el caso de las declaraciones conjuntas.
- Muestrear el conjunto

- De los individuos seleccionados se suministrarán sus **factores de elevación** y toda la información tal y como se ha establecido para el año base.
- Se buscará la existencia de **cónyuges en el conjunto total de individuos del año de referencia**, extrayéndose también la correspondiente información.
- Este colectivo se tratará como una subpoblación independiente, con sus correspondientes factores de elevación.

## ***Muestra en años anteriores al año base(1998-2002)***

- El muestreo en los años anteriores al año base será simétrico al definido para años posteriores. En una primera aproximación puede decirse que se trata de disponer de la información los años anteriores de los individuos seleccionados en el año base ( $t$ ).
- Para hacer el procedimiento correspondiente al año ( $t-1$ ) se deben conocer:
- Las altas habidas en el año 2003 (individuos de los que tenemos información en 2003 y no en 2002).

- Las bajas del año 2002 (los que están en el 2002 y no están en el 2003) que deberán muestrearse con los mismos criterios que los señalados para los años posteriores y utilizando la variable equivalente para la estratificación. Igualmente, deberá tratarse como una subpoblación independiente.
- Para el año 1998 se trataría únicamente de disponer de la información correspondiente a los individuos comunes.

- Las **bajas del año 2002** (los que **están en el 2002 y no están en el 2003**) que deberán muestrearse con los mismos criterios que los señalados para los años posteriores y utilizando la variable equivalente para la estratificación. Igualmente, deberá tratarse como una subpoblación independiente.
- Para el año 1998 se trataría únicamente de disponer de la información correspondiente a los individuos comunes.

# ESTIMADORES COMPLEJOS

Lo habitual es calcular los índices de Gini (IG) antes y después de la aplicación del impuesto

Calcular el índice de Reynolds-Smolensky (IRS), que expresa el grado de redistribución del impuesto así como la diferencia de los dos IG

El índice de Kakwani (1977) (IK), que mide la progresividad del impuesto mediante la diferencia entre el IG de la renta antes de impuestos y un índice de concentración de las cuotas líquidas ordenadas según renta. Todo está en función del índice de Gini

**El índice de Gini es una medida de concentración relativa definida como la mitad de la diferencia media para cada par de observaciones de renta, dividida por el valor medio de la variable cuya distribución se evalúa, tradicionalmente expresado como:**

$$G(y) = \frac{\sum_{i=1}^n \sum_{j=1}^n |y_i - y_j|}{2n^2 \bar{y}}$$

**Siguiendo a Glasser (1962) y Dixon (1987),  
alternativamente la fórmula del coeficiente de Gini  
puede escribirse como sigue:**

$$G(y) = \frac{1}{n(n-1)\bar{y}} \sum_{i=1}^n (2i - n - 1) y_i$$

**Adicionalmente, cuando se dispone de una muestra de tamaño n extraída de una población de tamaño N, el índice de Gini poblacional puede estimarse insesgadamente mediante el estimador siguiente:**

$$\hat{G}(y) = \frac{1}{N\bar{y}} \sum_{i=1}^n K_i y_i \left( 2 \sum_{j=1}^n K_j - K_i - N \right)$$

**donde y es la variable renta, n es el tamaño de la muestra,  $K_i$  es el factor de elevación y N es el tamaño poblacional**

# CUANTIFICACIÓN DEL ERROR AL ESTIMAR EL ÍNDICE DE GINI

Habitualmente el error absoluto de un estimador insesgado suele medirse, a partir de los datos de una muestra, mediante la estimación de su varianza.

Pero el problema aparece al intentar estimar la varianza cuando la expresión del estimador es complicada, tal y como ocurre en el caso del estimador del Índice de Gini.

En estas situaciones se acude a los métodos específicos de estimación de varianzas utilizados en la teoría del muestreo.

Entre estos métodos tenemos el método de las muestras interpenetrantes, el método de los grupos aleatorios, el método de las semimuestras reiteradas, el método de Jackknife y el método Bootstrap

# MUESTRAS INTERPENETRANTES

Se utiliza cuando tenemos un conjunto de dos o más muestras, elegidas con el mismo esquema de muestreo (independientes o no) y tales que cada una proporcione una estimación válida del parámetro que se pretenda estimar con el mismo error de muestreo.

Sean  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$  estimadores insesgados de  $\theta$  basados en  $k$  muestras independientes. Su media:

$$\hat{\theta} = \frac{1}{k} \sum_i^k \hat{\theta}_i$$

es también un estimador insesgado de  $\theta$

# SUBMUESTRAS INTERPENETRANTES

Se utiliza cuando tenemos un conjunto de dos o más muestras, elegidas con el mismo esquema de muestreo (independientes o no) y tales que cada una proporcione una estimación válida del parámetro que se pretenda estimar con el mismo error de muestreo.

Sean  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$  estimadores insesgados de  $\theta$  basados en  $k$  muestras independientes. Su media:

$$\hat{\theta} = \frac{1}{k} \sum_i^k \hat{\theta}_i$$

es también un estimador insesgado de  $\theta$

Un estimador insesgado para la varianza es :

$$\hat{V}(\hat{\theta}) = \frac{1}{k(k-1)} \left( \sum_i^k \hat{\theta}_i^2 - k\hat{\theta}^2 \right)$$

El estimador de la varianza para el índice de Gini será:

$$\hat{V}(\hat{G}) = \frac{1}{k(k-1)} \left( \sum_i^k \hat{G}_i^2 - k\hat{G}^2 \right)$$

$$\hat{G} = \frac{1}{k} \sum_i^k \hat{G}_i$$

En nuestro caso usamos 20 muestras independientes de tamaño 20.000 declaraciones del IRPF. El estimador de la varianza para el índice de Gini es:

$$\hat{V}(\hat{G}) = \frac{1}{k(k-1)} \left( \sum_i^k \hat{G}_i^2 - k\hat{G}^2 \right) = 0,00000293063$$

$$\hat{C}_v(\hat{G}) = \frac{\sqrt{\hat{V}(\hat{G})}}{\hat{G}} = 0,001295489$$

El error relativo para el estimador del índice de Gini es del 0,1295%, (uno por mil). Resultado óptimo derivado del elevado tamaño de las muestras, del elevado número de muestras y de su elevada precisión

# GRUPOS ALEATORIOS

Se extrae una muestra de  $n$  unidades de una población de tamaño  $N$ . Dicha muestra se subdivide en  $K$  submuestras de igual tamaño  $m$ , de modo que  $n=K.m$ . Estas submuestras se denominan grupos aleatorios, y además de ser submuestras de la muestra, también son muestras de la población completa.

En estas condiciones si  $\hat{\theta}$  es un estimador insesgado de la característica poblacional  $\theta$  basado en la muestra completa  $W$ , y si  $\hat{\theta}_r$  es un estimador insesgado de la característica poblacional  $\theta$  basado en el  $r$ -ésimo grupo aleatorio, un estimador insesgado de la varianza de es el siguiente:

$$\hat{V}(\hat{\theta}) = \frac{1}{K(K-1)} \sum_{r=1}^K (\hat{\theta}_r - \hat{\theta})^2$$

**En nuestro caso utilizamos 20 submuestras independientes de tamaño 20.000 declaraciones del IRPF. El estimador de la varianza para el índice de Gini será:**

$$\hat{V}(\hat{G}) = \frac{1}{K(K-1)} \sum_{r=1}^K (\hat{G}_r - \hat{G})^2 = 0,0000029792$$

$$\hat{C}_v(\hat{G}) = \frac{\sqrt{\hat{V}(\hat{G})}}{\hat{G}} = 0,001307335$$

# MÉTODO BOOTSTRAP (AUTOGENRACIÓN)

Para llevarlo a cabo partimos de la muestra de tamaño 400.000 declaraciones de IRPF extraída de una población de 16 millones de declaraciones. A continuación **extraemos de la muestra inicial M=1000 muestras con reposición**, también de tamaño un 400.000 y calculamos en cada una de ellas el estimador  $\hat{\theta}_j^*$  para el cual estamos calculando el error (índice de Gini). La precisión del estimador se obtiene por la expresión:

$$\hat{\sigma}_{BOOT\ 1000} = \sqrt{\frac{\sum_{j=1}^M (\hat{\theta}_j^*)^2 - \left(\sum_{j=1}^M (\hat{\theta}_j^*)\right)^2 / M}{M - 1}} = 0,00052414$$

$$\hat{C}_V(\hat{G}) = \frac{\hat{\sigma}_{BOOT1000}}{\hat{G}} = 0,00125543$$

**Si ahora consideramos M=5000 muestras, tenemos los siguientes resultados:**

$$\hat{\sigma}_{BOOT5000} = \sqrt{\frac{\sum_{j=1}^M (\hat{\theta}_j^*)^2 - \left( \sum_{j=1}^M (\hat{\theta}_j^*) \right)^2 / M}{M-1}} = 0,00052414$$

$$\hat{C}_V(\hat{G}) = \frac{\hat{\sigma}_{BOOT5000}}{\hat{G}} = 0,00122569$$

# INTERVALOS DE CONFIANZA NORMALIDAD

Método de las submuestras interpenetrantes.

El intervalo del índice de Gini es  $0,4175 \pm 0,001061$

Método de los grupos aleatorios: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,001069$

Método Bootstrap con 1000 muestras: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,001027$

Método Bootstrap con 5000 muestras: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,001003$

# NO NORMALIDAD

Método de las submuestras interpenetrantes.

El intervalo del índice de Gini es  $0,4175 \pm 0,00242$

Método de los grupos aleatorios: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,00244$

Método Bootstrap con 1000 muestras: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,00234$

Método Bootstrap con 5000 muestras: El intervalo de confianza para el índice de Gini es  $0,4175 \pm 0,00228$

# INTERVALOS DE CONFIANZA PARA EL ÍNDICE DE GINI

## Normalidad

$$\left[ \hat{\theta} - \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}}, \hat{\theta} + \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}} \right]$$

## No normalidad

$$\left[ \hat{\theta} - \lambda_{\alpha} \sigma(\hat{\theta}), \hat{\theta} + \lambda_{\alpha} \sigma(\hat{\theta}) \right]$$

# ALGUNAS CONCLUSIONES

Los errores obtenidos para las estimaciones del índice de Gini con la muestra y el panel son pequeños.

Este hecho indica la precisión de los cálculos de los índices de desigualdad, progresividad y redistribución basados en muestras y paneles.

De aquí se deriva la **pertinencia del uso de estos índices con fiabilidad en las simulaciones del impuesto basadas en las muestras y paneles del IRPF**

Por otro lado, se observa que los intervalos son más anchos cuando no hay normalidad (menos precisión en la estimación del índice de Gini).

Asímismo, se observa que el método menos afectado por la falta de normalidad en cuanto a precisión según los intervalos de confianza es el método Bootstrap.

Esto corrobora la idea de que el bootstrapping es una técnica muy adecuada como método especial para la estimación de varianzas en estimadores complejos.

- **IMPLICACIONES DE POLITICA FISCAL.**  
El panel permite la evaluación de reformas fiscales desde una perspectiva dinámica.
- En concreto, permite analizar el impacto de cualquier cambio impositivo en un mismo individuo (o unidad familiar o fiscal) a lo largo del tiempo. Es decir, permite un **análisis estructural de las reformas.**
- Esta es una cuestión fundamental en lo que a evaluación de reformas fiscales se refiere ya que **los individuos alteran su comportamiento en respuesta a los cambios impositivos.**

- El panel permite la evaluación de reformas fiscales desde una perspectiva dinámica.
- En concreto, permite analizar el impacto de cualquier cambio impositivo en un mismo individuo (o unidad familiar o fiscal) a lo largo del tiempo. Es decir, permite un **análisis estructural de las reformas**.
- Esta es una cuestión fundamental en lo que a evaluación de reformas fiscales se refiere ya que **los individuos alteran su comportamiento en respuesta a los cambios impositivos**.

- El panel permite la evaluación de reformas fiscales desde una perspectiva dinámica.
- En concreto, permite analizar el impacto de cualquier cambio impositivo en un mismo individuo (o unidad familiar o fiscal) a lo largo del tiempo. Es decir, permite un **análisis estructural de las reformas**.
- Esta es una cuestión fundamental en lo que a evaluación de reformas fiscales se refiere ya que los individuos alteran su comportamiento en respuesta a los cambios impositivos.

- Y muchos de los cambios de comportamiento no se producen en el instante posterior a la reforma sino uno o varios períodos después, por lo que la dimensión temporal del panel para introducir la dinámica es esencial.
- Los efectos de una reforma se pueden medir de modo más fiable cuando comparamos unidades homogéneas (individuos o unidades familiares o fiscales). Es esencial de trabajar con observaciones de los mismos individuos en distintos momentos del tiempo.

- Pero tampoco despreciamos las muestras anuales. Una sola muestra anual aislada puede no tener mucha fuerza, pero cuando se va construyendo un pool de muestras anuales sucesivas, la riqueza para el análisis aumenta exponencialmente e incluso permite contrastar los resultados obtenidos mediante análisis de panel.
- No olvidemos que las técnicas econométricas sobre panel no son fáciles de implementar, mientras que sobre pool de datos no hay tantas dificultades.

- Lo evidente es que, en el caso del IRPF, como se dispone de registros administrativos, el muestreo estratificado con criterios geográficos y algún otro tipo adicional de criterio (tramos de renta o fuentes de renta) es sencillo de realizar.
- Ello lleva a que sea muy adecuado disponer, tanto de sucesivas muestras anuales, como de un panel.
- A partir de las muestras anuales se pueden construir pool de datos.

# **PANEL DE DECLARANTES DE IRPF 1999-2007: METODOLOGÍA, ESTRUCTURA Y VARIABLES**

*Autores: Jorge Onrubia Fernández*  
Universidad Complutense de Madrid

*Fidel Picos Sánchez*  
Universidad de Vigo

*César Pérez López*  
Instituto de Estudios Fiscales

*Carmen Gallego Vieco*  
Instituto de Estudios Fiscales

*María del Carmen González Queija*  
Universidad de Vigo

*Sara Huete Vázquez*  
Instituto de Estudios Fiscales

DOC. n.º 7/2011

# **LIBRO BASE DEL PANEL**

**PANEL DE DECLARANTES DE IRPF**

**1999-2007**

**DISEÑO, METODOLOGÍA  
Y GUÍA DE UTILIZACIÓN**

Jorge Onrubia Fernández

Fidel Picos Sánchez

César Pérez López

**INSTITUTO DE ESTUDIOS FISCALES**

Universidad de Las Palmas de G. Canaria    Consejería de Economía y  
 Hacienda – Xunta de Galicia    Universidad Autónoma de Madrid  
 Universidad de la Laguna (Tenerife)    Fundación Acción Familiar  
 Universidad Autónoma de Barcelona  
 Universidad de Barcelona    Ayuntamiento de Barcelona    Universidad  
 de Extremadura  
 Universidad de Salamanca    Universidad de Valladolid    Dirección  
 General Economía – Baleares    Universidad de Navarra    Instituto  
 Cántabro de Estadística    Universidad Complutense de Madrid    Instituto  
 Cántabro de Estadística    Bolsa de Valores de Madrid  
 Fundación de Estudios de Economía Aplicada    Universidad de Valencia  
 Universidad de Granada    Instituto de Estadística de la Comunidad de  
 Madrid    Universidad de Zaragoza    Ecole Normale Supérieure    Universidad  
 California at Berkeley    Banco de España – Servicio de Estudios    Instituto  
 Nacional de Estadística

---

Fin de la presentación