



INSTITUTO NACIONAL DE ESTADÍSTICA



Prácticas encuestas estructurales económicas: Diseño de la Estadística Estructural de Empresas

*Departamento de Control de la Producción
Estadística y Muestreo
17 de junio de 2016*



Índice

1. Diseño de encuestas demográficas vs económicas
2. Diseño muestral de la Estadística Estructural de Empresas:
 - Estadística Estructural de Empresas: Industria, Comercio y Servicios. Diseño teórico
 - Marco muestral
 - Determinación de las empresas exhaustivas
 - Criterios de estratificación
 - Afijación muestral
 - Selección muestral: coordinación negativa
 - Estimadores
 - Errores de muestreo
 - Carga estadística e indicadores de calidad



1. Diseño de encuestas demográficas vs económicas

E. demográficas

- Marco de areas y de lista
- Muestreo multietápico
- Recogida de datos por entrevista personal
- Coste elevado
- Variables cualitativas
- Afijación de compromiso: Uniforme y proporcional
- Incidencias más comunes: Falta de respuesta

E. económicas

- Marco de lista
- Muestreo monoetápico
- Recogida multicanal. Proceso sancionador.
- Menor coste
- Variables cuantitativas. (Distribuciones asimétricas)
- Afijación óptima
- Incidencias más comunes: Unidades mal clasificadas



2. Diseño muestral de la Estadística Estructural de Empresas

- Estadística Estructural de Empresas: Industria, Comercio y Servicios. Diseño teórico.
- Marco muestral
- Determinación de las empresas exhaustivas
- Criterios de estratificación
- Afijación muestral
- Selección muestral: coordinación negativa
- Estimadores
- Errores de muestreo
- Carga estadística e indicadores de calidad



2. 1. Estadística Estructural de Empresas: Industria, Comercio y Servicios. Diseño teórico

- Encuesta Industrial Anual de Empresas
- Encuesta Anual de Comercio
- Encuesta Anual de Servicios

Estadística Estructural de Empresas:

- ☐ Sector Industrial
- ☐ Sector Comercio
- ☐ Sector Servicios



2. 1. Estadística Estructural de Empresas: Industria, Comercio y Servicios. Diseño teórico, cont...

Datos publicados (WEB):

Encuesta industrial de empresas - Año 2014

Número de empresas 188.470

Cifra de negocios 1 571.921.638

Personas ocupadas 1.931.972

La Encuesta industrial de empresas (EIAE) es una encuesta de carácter estructural, de periodicidad anual que tiene como objetivo fundamental proporcionar una información precisa, fiable y oportuna de los diversos sectores que constituyen la actividad industrial, de manera que se puedan satisfacer las necesidades de información en la materia, tanto nacionales como internacionales.

La encuesta se refiere al conjunto de empresas con actividad principal industrial (secciones B, C, D y E de la Clasificación Nacional de Actividades Económicas de 2009) y ubicadas en el territorio nacional.



2. 1. Estadística Estructural de Empresas: Industria, Comercio y Servicios. Diseño teórico, cont...

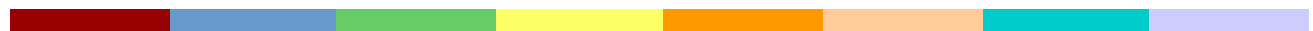
Variables estimadas:

- Personas ocupadas
- Horas trabajadas
- Ventas netas
- Prestaciones de servicios
- Cifra de negocios
- ...

Nivel de desagregación:

- Por Comunidades Autónomas
- Por agrupaciones de actividad

Peticiones a medida



2. 1. Estadística Estructural de Empresas: Industria, Comercio y Servicios. Diseño teórico, cont...

Muestreo aleatorio estratificado

- **Total de la característica Y:**

$$Y = \sum_{h=1}^L Y_h = \sum_{h=1}^L \sum_{i=1}^{N_h} y_{hi}$$

- **Estimador de expansión simple para Y:**

$$\hat{Y} = \sum_{h=1}^L \hat{Y}_h = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{y_{hi}}{\pi_i} = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{N_h}{n_h} y_{hi}$$

$$P(i \in s) = \pi_i = \frac{n_h}{N_h}$$

- **Peso de diseño=factor de elevación=** $d_h = \frac{1}{\pi_i} = \frac{N_h}{n_h}$



2. 1. Estadística Estructural de Empresas: Industria, Comercio y Servicios. Diseño teórico, cont.

- **Estimador de la varianza:**

$$\hat{V}(\hat{Y}) = \sum_{h=1}^L \hat{V}(\hat{Y}_h) = \sum_{h=1}^L \sum_{i=1}^{n_h} N_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}$$

donde

$$f_h = \frac{n_h}{N_h} \quad \text{fracción de muestreo}$$

$$\hat{S}_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}$$



2.2. Marco muestral

- DIRCE, marco anual.
- Información auxiliar:
 - Comunidad Autónoma
 - Código de actividad principal
 - Número de asalariados
 - Cifra de negocios
- **ACTIVIDAD PRINCIPAL** de la empresa es aquélla que genera el mayor valor añadido. Si no se dispone de esta información, se considera aquélla que proporcione el mayor valor de producción, o en su defecto, la que emplee un mayor número de personas ocupadas.



2.2. Marco muestral, cont...

Cada unidad se encuentra clasificada por su actividad principal, codificada según la Clasificación CNAE 2009.

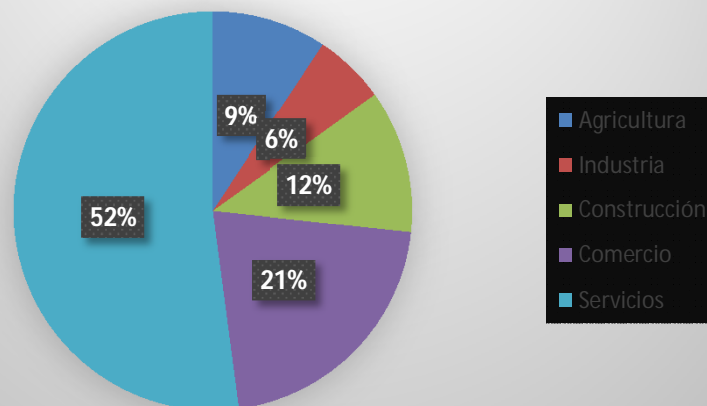
La CNAE-2009 está estructurada en cuatro niveles:

Nombre	Nivel	Nº de categorías	Identificación
Sección	Primero	21	Código alfabético de 1 letra
División	Segundo	88	Código numérico de 2 dígito
Grupo	Tercero	272	Código numérico de 3 dígito
Clase	Cuarto	629	Código numérico de 4 dígito

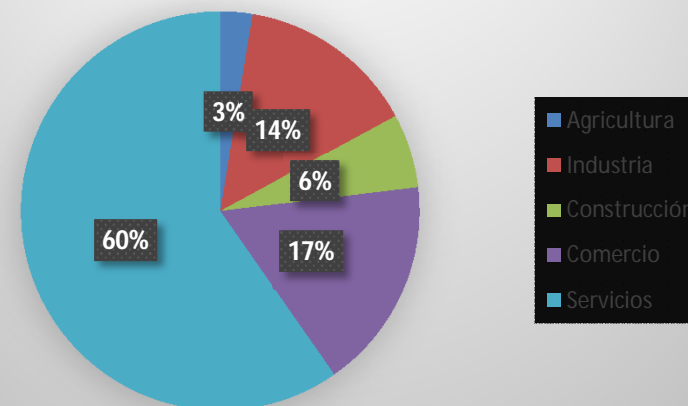


2. 2. Marco muestral, cont...

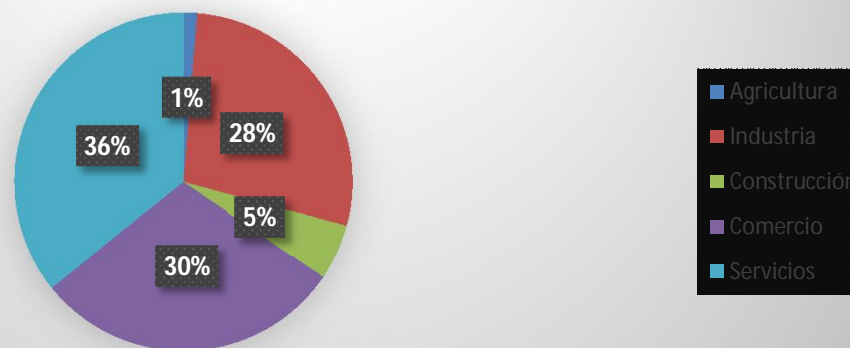
Porcentaje de empresas por actividad principal. Marco muestral 2014.



Porcentaje de asalariados por actividad principal. Marco muestral 2014.



Porcentaje de la cifra de negocios por actividad principal. Marco muestral 2014.



2. 2. Marco muestral, cont...

Datos del marco muestral 2014:

Número total de empresas	Distribución de empresas por grupos de asalariados, en %				
Año 2014	0 asalariados	De 1 a 9	De 10 a 49	De 50 a 199	Más de 199
3.173.178	52,9	42,2	4,1	0,6	0,2

Distribución de los asalariados, en %				
0 asalariados	De 1 a 9	De 10 a 49	De 50 a 199	Más de 199
0	24,5	19,8	13,8	41,9

Distribución de la cifra de negocios, en %				
0 asalariados	De 1 a 9	De 10 a 49	De 50 a 199	Más de 199
5,1	14,7	19,5	17	43,7



2. 2. Marco muestral, cont.

El marco muestral vigente, con el que se obtienen las muestras a recoger durante el 2016, hace referencia a la población de empresas del 2014:

- Encuestas de estructura: Referencia de los datos 2015
- Encuestas de coyuntura: Referencia de los datos 2016



2.3. Determinación de las empresas exhaustivas (autorepresentadas)

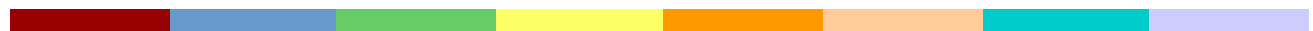
Búsqueda de los atípicos (outliers) por grandes, según cifra de negocios o facturación.

Objetivo:

- Pocas
- Relevantes
- Reduzcan la dispersión de la característica a investigar en la población restante

a) La regla de la desviación sigma.

b) El método del cuartil (Hidirogrou-Berthelot)



2.3. Determinación de las empresas exhaustivas, cont...

a) Regla de la desviación sigma: Sea U una población con N unidades y X una variable de la que disponemos de los valores x_1, x_2, \dots, x_N

- 1) Se ordenan las unidades, de menor a mayor, según una variable X:

$$x_{(1)} < x_{(2)} < \dots < x_{(N)}$$

- 2) Se calculan las diferencias $d_{(i)} = x_{(i)} - x_{(i-1)}$; $i=1 \dots N$

- 3) Se calcula la desviación típica de X:

$$\sigma(X) = \sqrt{V(X)} = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}}$$

- 4) Son exhaustivas todas las unidades que sigan a la primera en cumplir que la diferencia con la anterior, sea mayor que la desviación típica de dicha variable:

$$i_{(0)} / d_{(i_0)} > s(x)$$

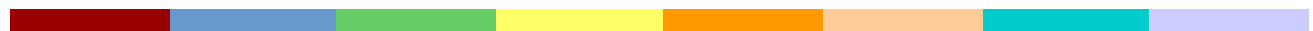
$i / i > i_0$ corresponden a unidades exhaustivas



2.3. Determinación de las empresas exhaustivas, cont...

a) Regla de la desviación sigma. Ejemplo.

- Grupo de actividad CNAE=3109 “Fabricación de muebles”
- Grupo de tamaño='1'=[1,3] asalariados
- $N=4.107$ empresas
 - Eliminamos las que no tienen X =Cifra de negocios y quedan:
 $N'=3.990$
 - Calculamos $\sigma(X)=245,21$
 - Ordenamos las empresas de menor a mayor según X :



CNAE=3109, TAME='1'		X=Cifra de negocios, miles de euros, ordenada de menor a mayor	
Nº de obs	X	Nº de obs	X
1	1	2501	98
2	1
....	...	3000	144
500	18
501	18	3982	1228
...		3983	1385
1000	34	3984	1833
1001	34	3985	2185
...	...	3986	2406
1500	51	3987	2436
1501	51	3988	2600
...	...	3989	3291
2500	98	3990	11737

2.3. Determinación de las empresas exhaustivas, cont...

a) Regla de la desviación sigma. Ejemplo.

- Calculamos las diferencias $d_i = x_i - x_{(i-1)}$ y comprobamos si $d_i > \sigma(X) = 245,21$
- En la obs 3984 ocurre $d_i = 1833 - 1385 = 448$
- Son exhaustivas a partir de la obs 3984. Total: 7 empresas exhaustivas.
- ❑ El coeficiente de variación de X, antes de separar las exhaustivas, es $CV(X) = 205,4\%$
- ❑ El coeficiente de variación de X, después de separar las exhaustivas, es $CV(X) = 118\%$



2.4. Determinación de las empresas exhaustivas, cont...

a) Regla de la desviación sigma. Aplicación en SAS.

-Aplicamos la regla en cada grupo de empresas formado por el cruce de clase de actividad (CNAE) y grupo de tamaño (TAME), medido en número de asalariados:

```
PROC SORT DATA=MARCO;BY CNAE TAME;
```

```
PROC UNIVARIATE DATA=MARCO NOPRINT;  
  VAR FACTU;  
  OUTPUT OUT=REGLA STD=STDF;  
  BY CNAE TAME;
```

```
PROC SORT DATA=MARCO;BY CNAE TAME FACTU;  
PROC SORT DATA=REGLA;BY CNAE TAME;
```



2.4. Determinación de las empresas exhaustivas, cont...

```
DATA REGLA1;  
  RETAIN AFACTU 0 ;  
  RETAIN CSFACTU 1.E+30 ;  
  MERGE MARCO REGLA; BY CNAE TAME;  
  
  IF FIRST.TAME THEN DO;  
    AFACTU=FACTU;  
    CSFACTU=1.E+30;  
  END;  
  ELSE IF FACTU-AFACTU GE STDF THEN DO ;  
    CSFACTU=FACTU;OUTPUT;  
  END ;  
  AFACTU=FACTU ;  
  IF LAST.TAME AND CSFACTU=1.E+30 THEN OUTPUT;  
  KEEP CNAE TAME CSFACTU ;
```



2.4. Determinación de las empresas exhaustivas, cont...

```
PROC SORT DATA=MARCO;BY CNAE TAME;  
PROC SORT DATA=REGLA1;BY CNAE TAME;
```

```
DATA EXHAUS1 QUEDA;  
MERGE REGLA1 MARCO;BY CNAE TAME;  
IF FACTU GE CSFACTU THEN OUTPUT EXHAUS1;  
ELSE OUTPUT QUEDA;  
RUN;
```

Inconvenientes de la regla de la desviación sigma:

- No siempre nos aseguramos de escoger las unidades más grandes o relevantes. Por ejemplo, en poblaciones donde X presenta una gran dispersión.
- Es posible seleccionar demasiadas unidades.



2.4. Determinación de las empresas exhaustivas, cont...

Soluciones:

- ☐ Complementarlo con otros métodos:
 - Un % de las más grandes según cada variable de interés en cada grupo o subpoblación.
 - **Método del cuartil**
- ☐ La selección de demasiadas unidades se evita poniendo un control en el número de unidades seleccionadas.



2.3. Determinación de las empresas exhaustivas, cont...

b) Método del cuartil

Se aplica por grupos de actividad a 4 dígitos y de tamaño, a la variable X =Cifra de negocios. Una empresa es exhaustiva si su cifra de negocios X_i supera al valor:

$$Me + C(Q_3 - Me)$$

Donde:

- Me es la mediana y Q_3 es el tercer cuartil de la distribución de X y
- C una constante, a fijar por el diseñador.



2.3. Determinación de las empresas exhaustivas, cont.

b) Método del cuartil. Aplicación en SAS.

```
PROC SORT DATA=QUEDA;BY CNAE TAME;
```

```
PROC UNIVARIATE DATA=QUEDA NOPRINT;
```

```
VAR FACTU;
```

```
OUTPUT OUT=QUEDA1 STD=STDF MEDIAN=MEFACTU Q3=Q3FACTU;  
BY CNAE TAME;
```

```
DATA QUEDA1;SET QUEDA1;
```

```
IF TAME IN('0','1','2') THEN C=9;
```

```
ELSE C=7;
```

```
LIMITES=MEFACTU+C*(Q3FACTU-MEFACTU);
```

```
PROC SORT DATA=QUEDA;BY CNAE TAME;
```

```
PROC SORT DATA=QUEDA1;BY CNAE TAME;
```

```
DATA EXHAUS2;
```

```
MERGE QUEDA QUEDA1;BY CNAE TAME;
```

```
IF FACTU GE LIMITES THEN OUTPUT;
```



2.4. Criterios de estratificación

Muestreo aleatorio estratificado

Objetivos:

- Estratos homogéneos (y estables)
- No demasiados estratos
- Evitar estratos con pocas unidades

Criterios a tener en cuenta:

- Dispersión de las variables a estimar y correlación con las variables auxiliares
- Precisión establecida por el Reglamento de la encuesta, Unidad Promotora o Muestreo
- Publicación de resultados



2.4. Criterios de estratificación, cont...

Estratos en el diseño antiguo:

Cruce de Comunidad Autónoma, clase de actividad y grupo de tamaño, medido en número de asalariados

CCAA×CLASE×TAMAÑO



2.4. Criterios de estratificación, cont.

Estratos en el nuevo diseño:

Objetivo: menos estratos y más grandes.

1 agrupación (MÍNIMO 10 empresas)	CCAA×CLASE×TAMAÑO
2 agrupación (MÍNIMO 10 empresas)	CCAA×SECTOR×TAMAÑO
3 agrupación (MÍNIMO 10 empresas excepto las empresas con 50 o más asalariados)	CLASE×TAMAÑO
4 agrupación	SECTOR×TAMAÑO



2. 5. Afijación muestral

Afijación óptima

Para unos errores relativos prefijados del estimador del total de número de asalariados o cifra de negocios, CV_0 , se calculan los tamaños muestrales n_h de forma que el tamaño muestral global (n) sea mínimo.

$$\begin{array}{l} \text{Mínimo} \\ \text{sujeto a} \end{array} \sum_h n_h$$

$$\frac{\sqrt{V_0(\hat{X})}}{X} \leq CV_0$$

Los errores relativos prefijados CV_0 (en la encuesta antigua y variable asalariados) son los siguientes:

- 1 % por clase de actividad
- 5 % por clase de actividad y región
- 20 % por clase de actividad, región y grupo de tamaño



2. 5. Afijación muestral, cont...

Afijación óptima

Solución:

$$n_h = \frac{N_h S_h (\sum_h N_h S_h)}{(CV_0 \times X)^2 + \sum_h N_h S_h^2}$$

Se calcula el n_h en cada uno de los tres casos:

$$CV_0 \times X_a = 0,01 \times \sum_{i \in \text{clase}} X_i$$

$$CV_0 \times X_{a,r} = 0,05 \times \sum_{i \in \text{clase}, \text{región}} X_i$$

$$CV_0 \times X_{a,r,t} = 0,20 \times \sum_{\substack{i \in \text{clase}, \text{región} \\ \text{tamaño}}} X_i$$



2. 5. Afijación muestral, cont...

$$n_{h,a} = \frac{N_h S_h (\sum_h N_h S_h)}{(CV_0 \times X_a)^2 + \sum_h N_h S_h^2}$$

$$n_{h,ar} = \frac{N_h S_h (\sum_h N_h S_h)}{(CV_0 \times X_{ar})^2 + \sum_h N_h S_h^2}$$

$$n_{h,art} = \frac{N_h S_h (\sum_h N_h S_h)}{(CV_0 \times X_{art})^2 + \sum_h N_h S_h^2}$$

$$CV_0 \times X_a = 0,01 \times \sum_{i \in \text{clase}} X_i$$

$$CV_0 \times X_{a,r} = 0,05 \times \sum_{i \in \text{clase, región}} X_i$$

$$CV_0 \times X_{a,r,t} = 0,20 \times \sum_{i \in \text{clase, región, tamaño}} X_i$$

$$n_h = \max(n_{h,a}, n_{h,ar}, n_{h,art})$$



2. 5. Afijación muestral, cont...

Afijación óptima en la nueva encuesta

CV₀ PREFIJADOS EN MUESTRA PRINCIPAL:

- SECTOR: 1%
- CCAA×DIVISIÓN: 5%
- Grupo de tamaño×DIVISIÓN: 5%

CV₀ PREFIJADOS EN MUESTRA TOTAL:

- CLASE: 1%
- CCAA×GRUPO: 10%
- Grupo de tamaño×GRUPO: 5%



2.6. Selección muestral: coordinación negativa

La coordinación muestral consiste en la manera de seleccionar muestras aleatorias, intentando controlar el grado de solapamiento entre ellas.

Coordinación muestral		
Positiva	MÁXIMO SOLAPAMIENTO	Muestras iguales o casi iguales
Negativa	MÍNIMO SOLAPAMIENTO	Muestras distintas o casi distintas



2.6. Selección muestral: coordinación negativa, cont...

Ventajas de las muestras coordinadas positivamente:

- Precisión (tasas de cambio)
- Depuración e imputación
- Coherencia
- Comparabilidad

Inconvenientes de las muestras coordinadas positivamente:

- ❑ Carga estadística
- ❑ Envejecimiento de la muestra
- ❑ Posibles sesgos de respuesta por la permanencia continua de la muestra



2.6. Selección muestral: coordinación negativa, cont...

Técnica de Números Aleatorios Permanentes (PRN, Permanent Random Number)

- A cada unidad del marco i ($i=1..N$) se le asocia un número aleatorio x_i obtenido de manera independiente de una distribución uniforme $U[0,1]$.
- Estos números aleatorios x_i , llamados PRNs se generan una vez y se guardan junto al resto de información del marco muestral, para usarlos en todas las extracciones muestrales que se realicen.



2.6. Selección muestral: coordinación negativa, cont...

PRN: Muestreo Aleatorio Simple, sin reemplazamiento

- i) Las unidades de la población se ordenan de manera ascendente(descendente) según el orden marcado por los x_i 's
- ii) Se eligen un punto de partida u origen de selección a_0 en $[0,1]$
- iii) Las 'n' unidades primeras de esta lista ordenada a partir del a_0 forman la muestra de tamaño 'n'



2.6. Selección muestral: coordinación negativa, cont...

PRN: Muestreo Aleatorio Simple, sin reemplazamiento. Aplicación en SAS

Obtener una muestra aleatoria estratificada de tamaño nh:

```
PROC SORT DATA=TAMANO;BY CCAA ACTI TAME;
PROC SORT DATA=MARCO;BY CCAA ACTI TAME;
DATA MARCO;
    MERGE TAMANO MARCO;BY CCAA ACTI TAME;
    ALEA=RANUNI(3);
PROC SORT DATA=MARCO;BY CCAA ACTI TAME ALEA;
DATA MUESTRA;SET MARCO;BY CCAA ACTI TAME;
    RETAIN CONTA 0;
    IF FIRST.TAME THEN CONTA=0;
    CONTA=CONTA+1;
    IF CONTA LE NH THEN OUTPUT;
```



2.6. Selección muestral: coordinación negativa, cont...

Controlar el solapamiento de dos muestras obtenidas de manera aleatoria, de tamaños n_1 y n_2

- i) Se ordenan las unidades del marco según los PRNs
- ii) Se eligen dos orígenes a_1 y a_2 en $[0,1]$
- iii) Para la primera muestra, se eligen las n_1 con los PRNs más cercanos a la derecha (o izquierda) de a_1
- iv) Para la segunda muestra se eligen los n_2 con los PRNs más cercanos a la derecha (o izquierda) de a_2



2.6. Selección muestral: coordinación negativa, cont...

Función de carga y PRNs anuales

$$f : \text{Marco} \rightarrow R^2$$

$$u_i \rightarrow f(u_i) = (x, y) =$$

(tiempo acumulado , número acumulado)



2.6. Selección muestral: coordinación negativa, cont...

Coordinación en un año determinado

- Inicialmente todas las unidades del marco tienen carga (0,0).
- La primera muestra se obtiene de manera independiente, utilizando un muestreo aleatorio simple dentro de cada estrato.
- Las unidades que han sido seleccionadas en la muestra pasan a tener una función de carga igual al tiempo necesario para cumplimentar el cuestionario y al número de encuestas que realiza la empresa, que en este caso es igual a 1.



2.6. Selección muestral: coordinación negativa, cont...

Coordinación en un año determinado

Primera encuesta: Tiempo para cumplimentar el cuestionario
=60'

$$f : \text{Marco} \rightarrow R^2$$

$$u_i \rightarrow f(u_i) = (x, y) = (\text{tiempo acumulado}, \text{número acumulado}) =$$
$$= \begin{cases} (60', 1) & \text{si } u_i \in s_1 \\ (0, 0) & \text{si } u_i \notin s_1 \end{cases}$$



2.6. Selección muestral: coordinación negativa, cont...

Coordinación en un año determinado

➤ Antes de obtener la siguiente muestra, se ordenan las unidades por orden creciente respecto:

- la primera componente de la función de carga (tiempo),
- la segunda componente (número de encuestas) y
- el número aleatorio

➤ Para las unidades seleccionadas en esta muestra, la función de carga se actualiza, acumulando el tiempo y el número de encuestas.



2.6. Selección muestral: coordinación negativa, cont...

Coordinación en un año determinado

Segunda encuesta: Tiempo para cumplimentar el cuestionario
=120'

$$f : \text{Marco} \rightarrow R^2$$

$$u_i \rightarrow f(u_i) = (x, y) = (\text{tiempo acumulado}, \text{número acumulado}) =$$

$$= \begin{cases} (180', 2) & \text{si } u_i \in s_1 \cap s_2 \\ (60', 1) & \text{si } u_i \in s_1 \\ (120', 1) & \text{si } u_i \in s_2 \\ (0, 0) & \text{si } u_i \notin s_1 \cup s_2 \end{cases}$$



2.6. Selección muestral: coordinación negativa, cont...

Aplicación en SAS

Obtener una muestra aleatoria estratificada de tamaño n_h , con función de carga:

```
PROC SORT DATA=TAMANO;BY CCAA ACTI TAME;  
PROC SORT DATA=MARCO;BY CCAA ACTI TAME;  
DATA MARCO;  
    MERGE TAMANO MARCO;BY CCAA ACTI TAME;  
    ALEA=RANUNI(3);  
PROC SORT DATA=MARCO;BY CAA ACTI TAME CARGA ALEA;  
DATA MUESTRA;SET MARCO;BY CCAA ACTI TAME;  
    RETAIN CONTA 0;  
    IF FIRST.TAME THEN CONTA=0;  
    CONTA=CONTA+1;  
    IF CONTA LE NH THEN OUTPUT;
```



2.7. Estimadores

- Estimador teórico del total de Y:

$$\hat{Y} = \sum_{h=1}^L \hat{Y}_h = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{N_h}{n_h} y_{hi}$$

- N_h : número total de empresas en el marco del estrato h
- n_h : número de empresas seleccionadas para la muestra en h
- y_{hi} : valor de la variable Y de la empresa i en h

Cada estrato h tiene asociado un único factor de elevación: N_h/n_h

Es un estimador insesgado $E(\hat{Y}) = Y$



2.7. Estimadores, cont...

Incidencias o estados de recogida:

- **Bajas, Erróneamente Incluidas, Duplicados:**
Reducen la población. (Exceso de cobertura)
- **Altas:** Incrementan la población. (Falta de cobertura)
- **Localizables, negativas:** Reducen la muestra teórica. (Falta de respuesta)
- **Cambios de estrato:** Redistribuyen la población entre los estratos. (Errores de clasificación)

Estimador AJUSTADO:

Estimadores de expansión simple ajustados por la falta de respuesta y los cambios de estrato.



2.7. Estimadores, cont...

Estimador AJUSTADO:

$$\hat{Y} = \sum_h \left(\sum_{i=1}^{n_h^*} \frac{\hat{N}_h^*}{n_h^*} y_{hi} + \sum_{k \neq h} \frac{N_k}{n_k} \sum_{i=1}^{n_k^h} y_{ki} \right)$$

- El primer sumando representa la aportación de las empresas que no han cambiado de estrato.
- El segundo sumando representa la aportación de las empresas seleccionadas en el estrato k y que realmente pertenecen al h .

Consecuencias de la falta de respuesta: Incremento del error muestral y aparición de sesgos.

Más de un factor de elevación por estrato.



2.7. Estimadores, cont...

Cálculo del factor de elevación final:

- Las unidades que cambian de estrato, mantienen su factor original:

$$\frac{N_h}{n_h}$$

- Las unidades que se mantienen en su estrato original, se reponderan. El factor de elevación queda:

$$\frac{\hat{N}_h^*}{n_h^*}$$



2.7. Estimadores, cont...

$$\frac{\hat{N}_h^*}{n_h^*}$$

- n_h^* : representa el número de empresas de la muestra efectiva (responden) que no han cambiado de estrato.
- \hat{N}_h^* : número de empresas estimado en el estrato h , obtenido al tener en cuenta las bajas y cambios de estratos.

$$\hat{N}_h^* = N_h \left(1 - \frac{b_h}{n_h}\right) - \sum_{k \neq h} \frac{N_h}{n_h} n_h^k$$

Donde n_h^k representa el número de unidades seleccionadas en el estrato h que son del estrato k y b_h número de bajas que reducen la población.



2.7. Estimadores, cont...

- Incidencias o estados de recogida

Incidencias y su significado		
E	Encuestada	Se ha recibido el cuestionario con datos
CP	Cierre parcial	Cierre estacional u ocasional
CD	Cierre definitivo	Baja confirmada
EI	Erróneamente Incluida	No corresponde al ámbito de esta encuesta
N	Negativa	De manera explícita, se niega a colaborar
IL	llocalizable	No se ha podido contactar con el informante
O	Otras	Otras incidencias no clasificadas anteriormente



■ Incidencias o estados de recogida

Encuesta Anual de Servicios 2013. Distribución de las empresas por tamaño e incidencias.

	Incidencias							Total
	E	CP	CD	EI	N	IL	O	
	Empresas %	Empresas %	Empresas %	Empresas %	Empresas %	Empresas %	Empresas %	Empresas %
Tamaño								
De 0	17956 57	922 3	1439 5	1082 3	1526 5	6615 21	2195 7	31735 100
De 1 a 2	17384 72	320 1	727 3	570 2	1323 5	3250 13	665 3	24239 100
De 3 a 5	11863 78	200 1	380 3	331 2	731 5	1396 9	249 2	15150 100
De 6 a 9	7089 80	160 2	249 3	189 2	399 4	674 8	114 1	8874 100
De 10 a 19	10408 82	314 2	362 3	302 2	431 3	741 6	162 1	12720 100
De 20 a 49	12323 83	417 3	619 4	311 2	505 3	580 4	171 1	14926 100
De 50 a 99	5354 84	184 3	268 4	126 2	162 3	227 4	59 1	6380 100
Más de 99	5187 91	85 1	192 3	97 2	34 1	80 1	18 0	5693 100
Total	87564 73	2602 2	4236 4	3008 3	5111 4	13563 11	3633 3	119717 100



2.8. Errores de muestreo

El error de muestreo relativo o coeficiente de variación:

Para un estimador insesgado $E(\hat{Y}) = Y$

$$CV(\hat{Y}) = 100 \times \frac{\sqrt{V(\hat{Y})}}{E(\hat{Y})} = 100 \times \frac{\sqrt{V(\hat{Y})}}{Y}$$

El error de muestreo relativo estimado:

$$C\hat{V}(\hat{Y}) = 100 \times \frac{\sqrt{\hat{V}(\hat{Y})}}{\hat{Y}}$$



2.8. Errores de muestreo, cont...

- Estimador del total

$$\hat{Y} = \sum_h \left(\sum_{i=1}^{n_h^*} \frac{\hat{N}_h^*}{n_h^*} y_{hi} + \sum_{k \neq h} \frac{N_k}{n_k} \sum_{i=1}^{n_k^h} y_{ki} \right)$$

Expresión del estimador de la varianza cuando se tienen en cuenta las incidencias (bajas, negativas, ilocalizables, cambios de estrato..)

$$\begin{aligned} \hat{V}(\hat{Y}) = & \sum_h \hat{N}_h^* (\hat{N}_h^* - n_h^*) \frac{\sum_{i=1}^{n_h^*} (y_{hi} - \hat{Y}_h^*)^2}{n_h^* (n_h^* - 1)} + \sum_h \hat{Y}_h^{*2} \hat{N}_h^* (N_h - \hat{N}_h^*) \frac{N_h - n_h}{N_h (n_h - 1)} + \\ & + \sum_{k \neq h} N_k (N_k - n_k) \frac{(s_k^h)^2}{n_k} \end{aligned}$$



2.8. Errores de muestreo, cont...

donde

$$\hat{Y}_h^* = \frac{\sum_{i=1}^{n_h^*} y_{hi}}{n_h^*} = \bar{y}_h^*; \quad (s_k^h)^2 = \frac{\sum_{i=1}^{n_k^h} y_{ki}^2}{n_k^h - 1} - \frac{(\sum_{i=1}^{n_k^h} y_{ki})^2}{n_k^h (n_k^h - 1)}$$

- Aproximación usando el método de Raulin

Sea el estimador del total en el dominio 'd' dado por:

$$\hat{Y}_d = \sum_h \sum_{i=1}^{m_h} F_{hi}' y_{hi} z_{hid} \quad \text{donde} \quad z_{hid} = \begin{cases} 1 & \text{si } i \in d \\ 0 & \text{en otro caso} \end{cases}$$



2.8. Errores de muestreo, cont.

- F'_{hi} es el peso o factor de elevación final asociado a la unidad i .
- y_{hi} es el valor que toma la variable Y en la unidad i del estrato h .
- m_h es el número de unidades que responden en h .

$$\hat{V}(\hat{Y}_d) = \sum_h \sum_{i=1}^{m_h} F'_{hi} (F'_{hi} - 1) (y_{hi} z_{hid} - \hat{Y}_{hd})^2 \quad \text{donde}$$

$$\hat{Y}_{hd} = \frac{\sum_{i=1}^{m_h} F'_{hi} y_{hi} z_{hid}}{\sum_{i=1}^{m_h} F'_{hi}} \quad y \quad m_h \approx m_h - 1$$



2.9. Carga estadística e indicadores de calidad

Carga estadística:

- Esfuerzo que realizan las unidades informantes al cumplimentar las encuestas.
- Se mide por el número de encuestas que realizan y el tiempo estimado para cumplimentarlas.
- Cada año, se elabora el Directorio de Colaboración de las Unidades Informantes, DICOIN:



2.9. Carga estadística e indicadores de calidad, cont...

DICOIN

- Para cada empresa, identificada por su NIF, se obtiene:
 - Número de encuestas económicas que realiza durante el año.
 - Incidencias o estados de recogida (responde, negativa, ilocalizable, cierre...)
 - El tiempo estimado para cumplimentarlas.
- La unidad de registro es el NIF.
- Cuando una encuesta va dirigida al establecimiento o centro de cotización, la información se agrega por NIF.



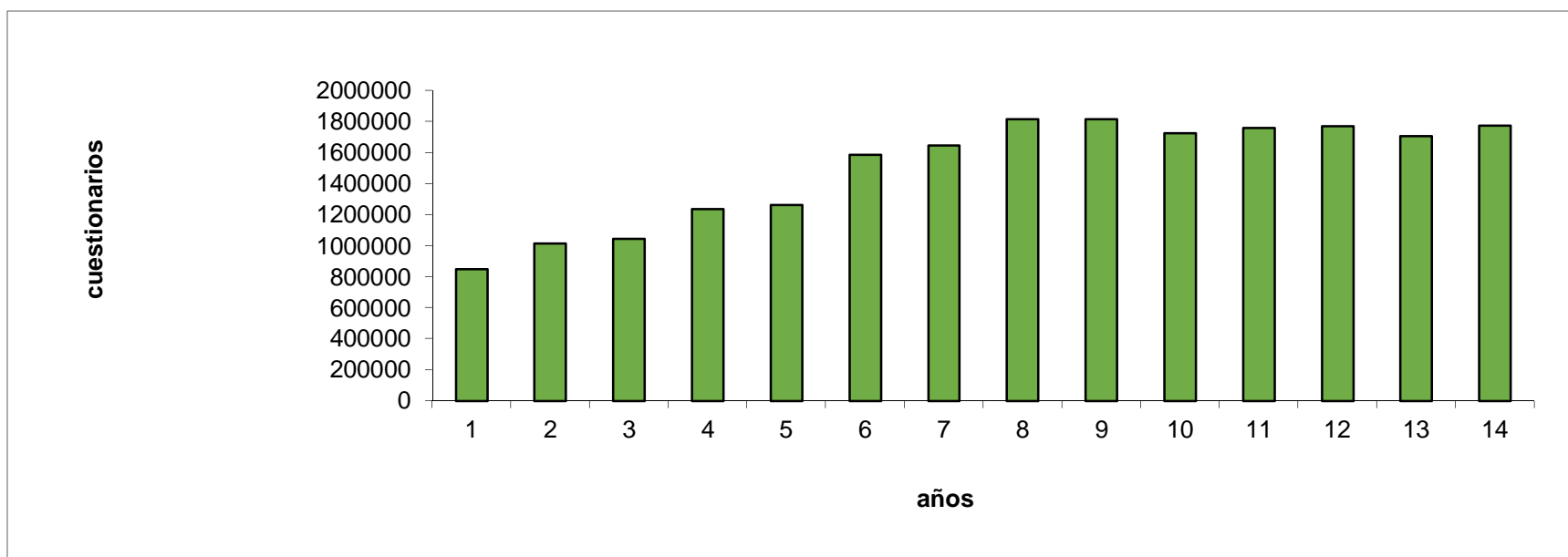
2.9. Carga estadística e indicadores de calidad, cont...

Las encuestas del DICOIN 14 son:

- 1 Encuesta anual de servicios
- 2 Encuesta industrial anual de empresas
- 3 Encuesta de consumos energéticos
- 4 Encuesta industrial anual de productos
- 5 Encuesta de innovación tecnológica
- 6 Encuesta de uso de TIC y comercio electrónico
- 7 Encuesta anual de coste laboral
- 8 Encuesta de estructura salarial
- 9 Encuestas de medio ambiente
- 10 Encuesta trimestral de coste laboral (ETCL)
- 11 Índice de producción industrial (IPI)
- 12 Índice de precios industriales (IPRI)
- 13 Índice de cifra de negocios y de entrada de pedidos en la industria
- 14 Índice de precios de importación y de exportación (IPRIX)
- 15 Índice de comercio al por menor (ICM)
- 16 Indicadores de actividad del sector servicios (IASS)
- 17 Encuesta trimestral de comercio internacional de servicios (CIS)
- 18 Indicadores de confianza empresarial (ICE)
- 19 Índice de precios del sector servicios (IPS)
- 20 Encuesta de ocupación hotelera
- 21 Encuesta de ocupación en apartamentos turísticos
- 22 Encuesta de ocupación en acampamentos turísticos
- 23 Encuesta de ocupación en alojamientos de turismo rural
- 24 Encuesta de ocupación en albergues



2.9. Carga estadística e indicadores de calidad, cont...



2.9. Carga estadística e indicadores de calidad, cont...

DICOIN 2014:

3.1. Distribución de las empresas por tamaño y número de encuestas en las que participa.
Porcentaje respecto al tamaño.

	TOTAL		TAMAÑO (NÚMERO DE ASALARIADOS)											
			De 0		De 1 a 9		De 10 a 19		De 20 a 49		De 50 a 99		Más de 99	
Número de encuestas	Empresas	%	Empresas	%	Empresas	%	Empresas	%	Empresas	%	Empresas	%	Empresas	%
1	230145	75	57900	98	120790	86	31745	63	16229	45	2162	19	1319	12
2	34344	11	938	2	9398	7	11512	23	8562	24	2634	23	1300	11
3	19315	6	245	0	7203	5	3494	7	4667	13	2234	20	1472	13
4	8764	3	34	0	1945	1	1771	4	2464	7	1148	10	1402	12
5	5383	2	12	0	401	0	1111	2	1936	5	775	7	1148	10
6	3264	1	2	0	66	0	356	1	1269	3	744	7	827	7
7	2122	1	2	0	9	0	120	0	644	2	614	5	733	6
8	1561	1	1	0	8	0	48	0	280	1	463	4	761	7
9	1116	0	21	0	142	0	263	2	690	6
10	793	0	6	0	87	0	153	1	547	5
11	485	0	33	0	104	1	348	3
12	355	0	5	0	68	1	282	2
13	297	0	2	0	19	0	276	2
14	263	0	2	0	261	2
15	79	0	79	1
16	1	0	1	0
TOTAL	308287	100	59134	100	139820	100	50184	100	36320	100	11383	100	11446	100

2.9. Carga estadística e indicadores de calidad, cont...

DICOIN 2014:

Distribución de las empresas por tamaño y tiempo estimado en cumplimentar el cuestionario. Porcentaje respecto al tamaño. Año 2013.														
	Total		Tamaño (Número de Asalariados)											
			De 0		De 1 a 9		De 10 a 19		De 20 a 49		De 50 a 99		Más de 99	
	Empresas	%	Empresas	%	Empresas	%	Empresas	%	Empresas	%	Empresas	%	Empresas	%
Tiempo (Horas)														
<=1	51376	21	9817	29	28958	26	9361	22	2891	9	213	2	136	1
(1,2]	13902	6	3295	10	6146	6	2768	7	1420	4	152	1	121	1
(2,3]	99150	41	17645	51	54620	49	13476	32	10370	32	2096	20	943	9
(3,4]	27443	11	1862	5	9394	8	8796	21	5215	16	1295	12	881	8
(4,5]	6312	3	1235	4	2128	2	1078	3	1319	4	309	3	243	2
(5,10]	26138	11	416	1	8514	8	4344	10	6550	20	3483	33	2831	26
(10,20]	13130	5	54	0	1278	1	1980	5	4433	14	2468	23	2917	27
> 20 horas	3959	2	10	0	18	0	61	0	408	1	676	6	2786	26
Total	241410	100	34334	100	111056	100	41864	100	32606	100	10692	100	10858	100



2.9. Carga estadística e indicadores de calidad, cont...

- Acortamiento de los cuestionarios
- Uso de datos administrativos
- Depuración Selectiva
- Selección coordinada de muestras teniendo en cuenta la carga estadística

Compensar la carga estadística:

Ofrecer a las empresas información gratuita, hecha a medida, que le pueda interesar:

- Ratios del sector
- Cuota de mercado
- Análisis de la estructura del sector



2.9. Carga estadística e indicadores de calidad, cont...

Indicadores de calidad

- Errores de muestreo
- Errores ajenos al muestreo

Indicadores:

- Coeficiente de variación.
- Tasa de respuesta por unidad, ponderada y sin ponderar.
- Tasa de respuesta por ítem, ponderada y sin ponderar.
- Tasa de imputación y proporción imputada.
- Tasas de exceso de cobertura y de clasificación errónea.
- Tasa de falta de cobertura.



2.9. Carga estadística e indicadores de calidad, cont...

➤ Tasa de respuesta, sin ponderar:

$$TR = 100 \times \frac{E}{T - CD - EI}$$

Donde:

T=Total de unidades en la muestra

CD=Cierres definitivos

EI=Erróneamente Incluidas



2.9. Carga estadística e indicadores de calidad, cont...

➤ Tasa de respuesta, ponderada:

$$TR_p = 100 \times \frac{\sum_{h=1}^L E_h \times \frac{N_h}{n_h}}{\sum_{h=1}^L N_h - \sum_{h=1}^L CD_h \times \frac{N_h}{n_h} - \sum_{h=1}^L EI_h \times \frac{N_h}{n_h}}$$

Donde:

N_h =Número de unidades en la población inicial del estrato h

n_h =Número de unidades en la muestra inicial en el estrato h



2.9. Carga estadística e indicadores de calidad, cont...

7.2. Tasa no ponderada de respuesta y de no-respuesta. Año 2013

			Tasa de respuesta %	Tasa de no respuesta %				
			Total	Total	Motivos de no-respuesta			
					CP	N	IL	O
E. Anual de Servicios			78	22	2	5	12	3
E. Industrial Anual de Empresas			87	13	3	4	6	0
E. De Innovación Tecnológica			87	13	3	1	6	3
E. De Uso del TIC y Comercio Electrónico			81	19	6	0	10	3



2.9. Carga estadística e indicadores de calidad, cont...

Errores de cobertura:

Los errores de cobertura se deben a diferencias entre la población objetivo y el marco e incluyen:

- Exceso de cobertura: Unidades del marco que no forman parte de la población objetivo.
- Falta de cobertura: Unidades de la población objetivo que no están en el marco.
- Clasificación errónea: Unidades del marco con información auxiliar, usada para estratificar, diferente a la que realmente tienen en la población objetivo.



2.9. Carga estadística e indicadores de calidad, cont...

Índice de Exceso de Cobertura (IEC):

Porcentaje entre los Cierres Definitivos (CD) más las Erróneamente Incluidas (EI) y el Total de unidades muestrales (T), exceptuando las Ilocalizable (IL) y otras (O).

$$IEC = 100 \times \frac{CD + EI}{T - IL - O}$$



2.9. Carga estadística e indicadores de calidad, cont...

Índice de clasificación errónea o cambio de clasificación (ICC):

Porcentaje entre las unidades que cambian de clasificación o estrato (CC) y las Encuestadas (E) más los Cierres Parciales (CP) y las Negativas (N). Se calcula respecto a las principales variables de estratificación: Actividad y Tamaño.

$$ICC = 100 \times \frac{CC}{E + CP + N}$$



2.9. Carga estadística e indicadores de calidad, cont.

7.1. Indicadores de calidad para Errores de Cobertura. Año 2013.

			IEC %	ICC %	
				Estrato de tamaño	Estrato de actividad
E. Anual de Servicios			7	26	11
E. Industrial Anual de Empresas			9	16	2
E. De Innovación Tecnológica			10	17	7
E. De Uso del TIC y Comercio Electrónico			4	12	5

