Curso Selectivo 2016

Diseño muestral en las encuestas del INE y evaluación de la calidad de los datos

junio 2016

Introducción

INSTITUTO NACIONAL DE ESTADISTICA

En su actividad estadística el INE realiza, entre otras, dos grandes tipos de operaciones:

- **Censos**: Investigaciones de tipo exhaustivo.
  - Censo de Población y Vivienda. 2001, 2011
  - Censo Agrario. 1999, 2009
  - Censo Industrial (actualmente no se realiza)

- **Encuestas por muestreo**: Proceso mediante el cual se obtienen estimaciones de parámetros de la población a partir de la información proporcionada por una parte de ella (*muestra*).

INSTITUTO NACIONAL DE ESTADISTICA

Introducción

Censos

- Muy costosos
- Permiten una mayor desagregación geográfica
- Pocas variables objetivo
- Sin errores de muestreo
- Errores ajenos al muestreo

Encuestas

- Menos costosas /Más frecuentes
- Problemas con la estimación en áreas pequeñas
- Más variables objeto de estudio
- Con errores de muestreo
- Con menos errores ajenos al muestreo

Introducción

Complementariedad entre Censos y Encuestas

Los Censos proporcionan a las encuestas por muestreo información necesaria para:

- Preparación de los **marcos** de muestreo
- Definición de **estratos y subestratos**
- Mejora de los procesos de **estimación**

Principales etapas de una encuesta

INSTITUTO NACIONAL DE ESTADISTICA

INSTITUTO NACIONAL DE ESTADISTICA

Principales etapas de una encuesta

Tipos de encuestas realizadas en el INE

INSTITUTO NACIONAL DE ESTADISTICA

Según el *ámbito poblacional*

- Encuestas de **Población**: Dirigidas a los hogares.
- Encuestas **Económicas**: Dirigidas a las empresas.

Según la *periodicidad*

- Encuestas **Continuas**
- Encuestas **Esporádicas**

Según el *tipo de la información* que proporciona

- Encuestas **Estructurales**
- Encuestas **Coyunturales**

Tipos de encuestas realizadas en el INE

Encuestas de población y hogares

INSTITUTO NACIONAL DE ESTADISTICA

Encuestas continuas

- Encuesta de Población Activa
- Encuesta Continua de Presupuestos Familiares
- Encuesta de Condiciones de Vida
- Encuesta de uso de TIC en los Hogares
- Encuesta Continua de Hogares

Encuestas esporádicas

- Encuesta de Empleo del Tiempo
- Encuesta de Discapacidades, Autonomía Personal y Situaciones de Dependencia
- Encuesta Nacional / Europea de Salud
- Encuesta de Inserción Laboral de los Titulados Universitarios

Tipos de encuestas realizadas en el INE

INSTITUTO NACIONAL DE ESTADISTICA

Tamaño de la muestra

Encuesta	Per.	U.P.		U.S.	
(Hogares)					
E. de Población Activa	Trim.	Sec	**3.588**	Viv.	**70.000**
E.C.de Presupuestos Familiares	Anual	Sec	**2.054**	Viv.	**22.000**
E. Cond.de Vida(EU-SILC)	Anual	Sec	**2.000**	Viv.	**16.000**
TIC_Hogares	Anual	Sec	**2.580**	Viv.	**21.000**
E.Nacional de Salud(2011)		Sec	**2.000**	Viv.	**24.000**
E.Europea de Salud(2014)		Sec	**2.500**	Viv.	**37.500**
E.Inserción Laboral Univ.(2014)		Tit.	**40.000**		
E.Empleo Tiempo (2009)		Sec	**1.232**	Viv.	**11.182**

Tipos de encuestas realizadas en el INE

Encuestas económicas (empresas)

INSTITUTO NACIONAL DE ESTADISTICA

Encuestas estructurales:

- Encuesta Anual de Servicios
- Encuesta Industrial Anual de Empresas
- Encuesta de Innovación Tecnológica
- Encuesta de uso del TIC y Comercio Electrónico
- Encuestas medioambientales
- Encuesta Industrial de Productos
- Encuesta sobre la Estructura de las Explotaciones Agrícolas

Encuestas coyunturales

- Índices coyunturales (ICM,IASS,ETCL)
- Encuesta mensual de transportes de viajeros
- Encuestas de Turismo

Tipos de encuestas realizadas en el INE

INSTITUTO NACIONAL DE ESTADISTICA

Tamaño de la muestra

Encuesta	**Per.**	**U. elementales**	
(Empresas)			
E. Industrial Anual(EIA)	Anual	Empresas	**45.000**
E.Servivcios(EAS)	Anual	Empresas	**120.000**
Innovación tecnológica	Anual	Empresas	**25.000**
I+D(exhaustiva)	Anual	Empresas	**18.000**
Estructura Explotaciones Agrícolas	Triena	Explot.	**65.000**
Coste Laboral(ETCL)	Trim.	Centros	**28.000**
E. TIC Empresas	Anual	Empresas	**28.000**
Índice de Confianza Empresarial	Trim.	Estab.	**7.500**

Introducción al diseño muestral

INSTITUTO NACIONAL DE ESTADISTICA

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

Información previa necesaria

Variable objetivo

Tipo de estimaciones

Precisión de las estimaciones

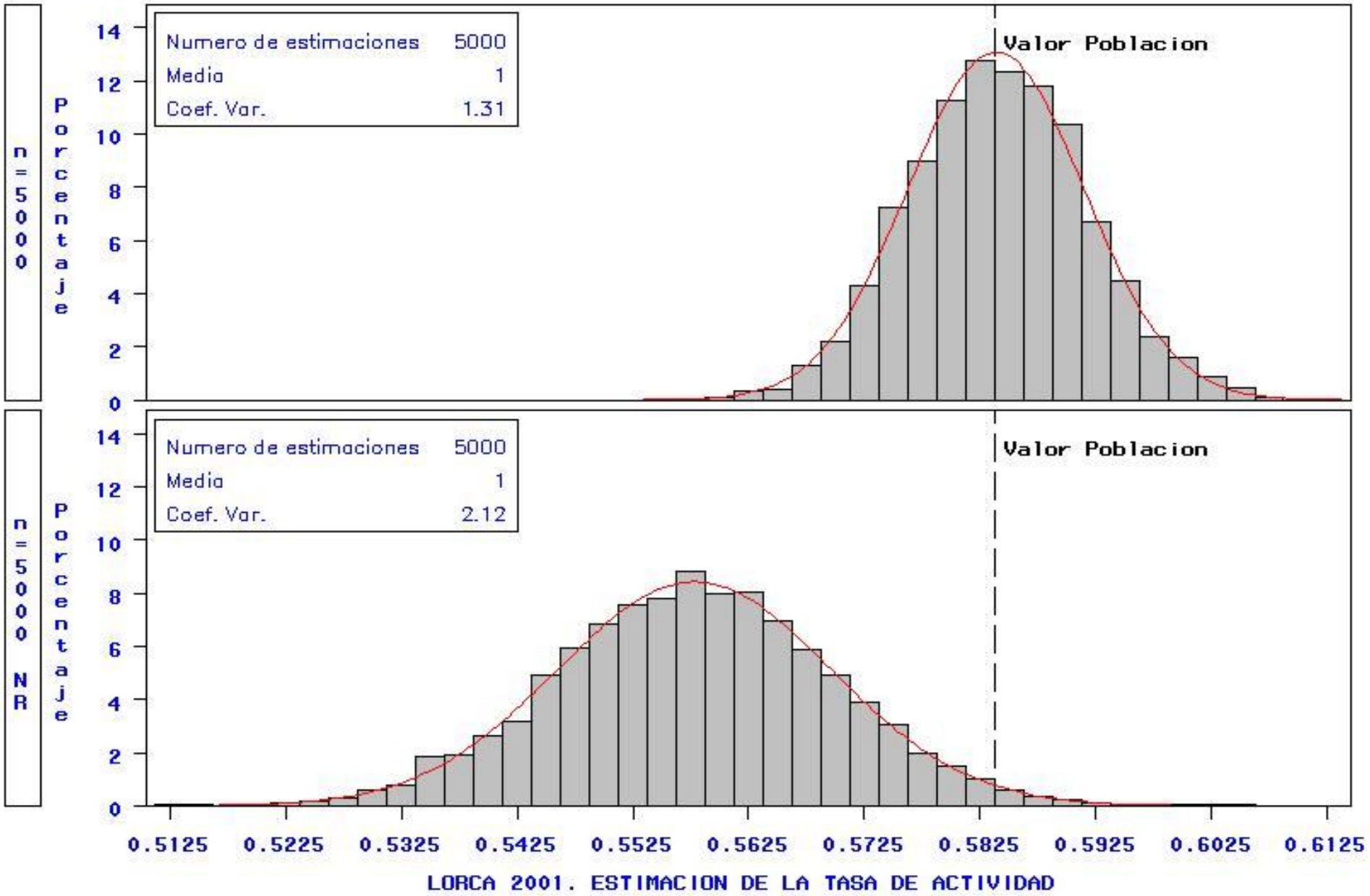
Tablas

Experiencias anteriores

Introducción al diseño muestral

Precisión.- La precisión de una estimación se mide por el error cuadrático medio (ECM), que incluye:

- **Sesgo.** Diferencia entre el valor esperado de la estimación y el valor poblacional. Normalmente se utilizan estimadores insesgados (o cuasi), de modo que el sesgo de las estimaciones sería nulo de no ser por la presencia de errores ajenos al muestreo
- **Variabilidad.** Medida de la dispersión de las estimaciones. Depende *fundamentalmente* de:
  - La dispersión de la variable objetivo
  - El tamaño de la muestra (y en menor medida, del tamaño poblacional)
  - El estimador utilizado


n=5000
Porcentaje
Numero de estimaciones 5000
Media 1
Coef. Var. 1.31
Valor Poblacion
n=5000 NR
Porcentaje
Numero de estimaciones 5000
Media 1
Coef. Var. 2.12
Valor Poblacion
0 2 4 6 8 10 12 14
0.5125 0.5225 0.5325 0.5425 0.5525 0.5625 0.5725 0.5825 0.5925 0.6025 0.6125
LORCA 2001. ESTIMACION DE LA TASA DE ACTIVIDAD

Tipos de diseños muestrales

Encuestas continuas

Panel fijo

Ventajas: Proporciona estimaciones longitudinales

Desventajas: Es caro y complejo realizar el seguimiento temporal de las unidades muestrales. No proporciona estimaciones transversales.

Panel rotante: *Periódicamente se renueva una parte de la muestra. Es muy utilizado en los institutos de estadística*

*Ventajas: P*ermite obtener estimaciones del cambio y estimaciones transversales

Desventajas: La actualización de la muestra se produce de forma lenta, lo que dificulta la estimación de fenómenos muy dinámicos

Tipos de muestreo

INSTITUTO NACIONAL DE ESTADISTICA

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

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

Tipos de muestreo

Muestreo aleatorio simple (m.a.s.): *Selección de unidades elementales con probabilidades iguales*

Ventajas

Marco sencillo, sin necesidad de información auxiliar

Fácil programación de la selección de la muestra, los factores de elevación y los errores de muestreo

Cálculo sencillo del tamaño de la muestra necesario

Dispersión de la muestra

Desventajas

Coste elevado de los trabajos de campo por la dispersión muestral

La no utilización de información auxiliar impide mejorar la representatividad de la muestra

Tipos de muestreo

Muestreo estratificado. *Selección de muestras independientes en estratos, agrupaciones homogéneas de unidades elementales o dominios de estudio. En cada uno de los estratos se puede realizar el tipo de muestreo que se considere más adecuado*

***Ventajas**:*

Mejora la representatividad de la muestra, en lo que se refiere a las variables utilizadas en la estratificación

Mejora la precisión de las estimaciones globales

Permite garantizar una precisión dada en dominios de estimación

Permite un reparto o afijación muestral por estratos óptima en cuanto a la precisión de la estimación global y al coste del campo

Desventajas:

Es necesaria información auxiliar en el marco

Tipos de muestreo

Muestreo sistemático: *Selección periódica secuencial de una muestra, previa ordenación del marco*

Ventajas:

Implícitamente realiza una estratificación del marco, definida según los valores de las variables utilizadas en la ordenación previa

En encuestas de hogares, la ordenación previa de los mismos por calles proporciona una muestra bien distribuida geográficamente

Permite seleccionar muestras con probabilidades iguales, o con probabilidades proporcionales al tamaño de las unidades elementales

Desventajas:

Implícitamente realiza un muestreo por conglomerados

Posible sobre o subestimación en las estimaciones debida a periodicidades en la ordenación previa.

Tipos de muestreo

Muestreo por conglomerados sin submuestreo: *Selección de conglomerados, agrupaciones de unidades elementales (empresas, viviendas o secciones censales)*

Ventajas:

Facilita la elaboración del marco

Disminuye el coste de los trabajos de campo

Desventajas

Normalmente disminuye la precisión de las estimaciones debido al llamado efecto conglomerado (DEFF)

Tipos de muestreo

Muestreo por conglomerados con submuestreo: *Selección de conglomerados en primera etapa, y de unidades elementales dentro de cada conglomerado elegido, en segunda etapa*

Ventajas:

Reduce el coste en desplazamientos de los entrevistadores

Facilita el acceso del entrevistador a hogares y personas, al resultar conocida su persona en los barrios seleccionados

Simplifica la organización de los trabajos de campo

Desventajas:

El efecto conglomerado (DEFF)

Escasa dispersión de la muestra, que puede impedir la obtención de estimaciones directas en áreas pequeñas

Tipos de muestreo

Muestreo por cuotas: *Selección de una muestra equilibrada (cuotas), según variables de distribución conocida (p.ej. grupos de edad y sexo*

Se diferencia del muestreo estratificado en que una vez determinada la cuota, el investigador es libre de elegir a los sujetos de la muestra dentro de cada estrato.

Muestreo por rutas aleatorias: *Normalmente asociado al muestreo por cuotas, realiza la selección de la muestra mediante itinerarios aleatorios predefinidos*

Con ello evita la necesidad de un marco, pero no garantiza la aleatoreidad de la selección.

Esquema del diseño muestral de una encuesta

1. Ámbito
2. Marco de selección
3. Tipo de muestreo
4. Criterios de estratificación
5. Tamaño y afijación de la muestra
6. Distribución de la muestra en el tiempo
7. Selección
8. Renovación parcial de la muestra
9. Estimadores
10. Calidad de las estimaciones

Diseño muestral de encuestas demográficas

INSTITUTO NACIONAL DE ESTADISTICA

Ámbito

-**Poblacional**: Población que reside en viviendas familiares principales. Se excluyen los hogares colectivos.

-**Geográfico**: Territorio nacional.

-**Temporal**: *Resultados de la encuesta*. (EPA:Trimestre).

Información recogida: Semana anterior a la de la entrevista (semana de referencia)

Diseño muestral de encuestas demográficas

Marco

Relación de unidades que van a ser muestreadas junto con la información complementaria disponible

Como el muestreo es bietápico, se utilizan dos marcos

Marco de áreas geográficas (U.P.)

Secciones censales (UP). Aproximadamente 35.000

Marco de viviendas (U.S.)

Viviendas familiares principales con sus direcciones postales, en cada una de las secciones censales seleccionadas para la encuesta

Censo 2001

INSTITUTO NACIONAL DE ESTADÍSTICA

CPRO	CMUN	DIST	NSECC	Población	% de jóvenes (0-19)	% de jóvenes (15-24)	% de Mayores	% de parados en la sección	% de inactivos	% de ocupados	% de extranjeros
41	091	01	022	1.146,0	9,34	21,29	20,24	10,38	53,66	35,95	3,14
41	091	01	023	1.487,0	9,75	21,52	16,75	11,97	49,83	37,26	2,69
41	091	01	024	1.261,0	10,55	17,76	20,38	10,47	54,48	34,66	2,38
41	091	01	025	2.036,0	11,25	19,40	17,58	9,48	49,85	37,28	2,65
41	091	01	027	1.391,0	9,99	22,00	21,21	5,97	54,57	39,47	1,01
41	091	01	028	773,0	12,55	20,83	17,21	11,25	52,65	34,67	2,85
41	091	01	029	1.915,0	9,92	23,86	13,68	11,96	47,42	35,67	1,04
41	091	01	030	762,0	8,27	23,23	22,18	6,96	53,67	37,53	0,79
41	091	01	031	758,0	8,84	17,81	26,65	10,16	56,20	33,64	1,72

				% de personas con nivel de estudios							
CPRO	CMUN	DIST	Nº SECC	inferiores	medios	superiores	Renta total por vivienda con percentores	Renta por desempleo entre renta total	Renta Capital mobiliario e inmobiliario sobre renta total	Renta agraria sobre renta total	Subestrato
41	091	01	022	39,70	37,87	22,43	19160,6	2,0	4,6	0,1	4
41	091	01	023	36,85	42,57	19,64	17464,7	2,2	3,1	0,0	4
41	091	01	024	44,96	33,23	21,41	19662,2	1,6	5,2	0,3	4
41	091	01	025	43,71	34,33	18,57	18711,8	1,8	3,7	0,3	4
41	091	01	027	26,82	37,60	35,59	44987,0	0,5	23,4	1,2	6
41	091	01	028	46,18	33,12	19,28	19579,7	1,5	4,6	0,4	4
41	091	01	029	37,08	39,95	18,02	19480,2	1,7	4,6	0,3	4
41	091	01	030	29,27	43,96	24,93	33633,7	1,2	6,6	0,0	4
41	091	01	031	41,29	39,58	19,13	17857,5	2,7	4,1	0,1	4

Diseño muestral de encuestas demográficas

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

Unidades primarias: Secciones censales
Unidades secundarias: Viviendas familiares principales

Estratificación de unidades primarias

Criterio geográfico. Los *estratos* se definen en función del tamaño del municipio al que pertenece la sección

Criterio socioeconómico. Los *subestratos* se forman a partir de la información censal disponible, aplicando análisis de conglomerados

Diseño muestral de encuestas demográficas

Tamaño de la muestra

Se establece en función de:

1- La desagregación requerida para las estimaciones

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

3- Límites de precisión establecidos por el Servicio Promotor

4- Experiencia de otras encuestas anteriores o similares

5- *Presupuesto*

De acuerdo con lo anterior se determina:

- El número de secciones muestrales por estrato
- Un número fijo (o no) de viviendas por sección

Diseño muestral de encuestas demográficas

INSTITUTO NACIONAL DE ESTADISTICA

Tamaño de la muestra

Encuesta	**Per.**	**U.P.**		**U.S.**	
(Hogares)					
E. de Población Activa	Trim.	Sec	**3.588**	Viv.	**70.000**
E.C.de Presupuestos Familiares	Anual	Sec	**2.054**	Viv.	**22.000**
E. Cond.de Vida(EU-SILC)	Anual	Sec	**2.000**	Viv.	**16.000**
TIC_Hogares	Sem.	Sec	**2.580**	Viv.	**21.000**
E.Nacional de Salud(2011)		Sec	**2.000**	Viv.	**24.000**
E.Europea de Salud(2014)		Sec	**2.500**	Viv.	**37.500**
E.Inserción Laboral Univ.(2014)		Tit.	**40.000**		
E.Empleo Tiempo (2009)		Sec	**1.232**	Viv.	**11.182**

Diseño muestral de encuestas demográficas

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

+EPA Muestra de secciones		SEMANA												
MÁLAGA	ESTRATO	01	02	03	04	05	06	07	08	09	10	11	12	13
	1	3	3	3	2	3	3	2	2	3	3	3	3	3
	4	1	1	1	.	1	1	.	1	.	1	1	1	1
	5	2	1	1	2	1	1	2	2	2	1	1	1	1
	6	.	.	.	1	1	.	1	.	1	.	.	.	1
	7	.	1	1	1	.	1	1	1	.	1	1	1	.
	All	6	6	6	6	6	6	6	6	6	6	6	6	6

Diseño muestral de encuestas demográficas

Selección de la muestra

Secciones: Probabilidad proporcional al tamaño medido por el número de viviendas familiares principales

Viviendas: Probabilidad igual(muestreo sistemático)

De esta forma en cada estrato, las viviendas familiares tienen la misma probabilidad de pertenecer a la muestra (**muestra autoponderada**)

$$P(V_{ijh}) = P(S_{jh}).P(V_{ijh}/S_{jh}) = K_h.\frac{V_{jh}}{V_h}.\frac{m}{V_{jh}} = \frac{K_h.m}{V_h} = \frac{v_h^t}{V_h}$$

Diseño muestral de encuestas demográficas

Renovación parcial de la muestra

Unidades primarias: Las secciones censales. Permanecen fijas indefinidamente en la muestra (salvo agotamiento de los hogares consultables o actualización de probabilidades de selección).

Unidades secundarias: Las viviendas familiares. Son renovadas parcialmente cada trimestre. En EPA esta renovación afecta a una sexta parte de las secciones (5/6 permanecen de un trimestre a otro)

Turnos de rotación: El conjunto de las secciones de la muestra está repartido en 6 grupos llamados turnos de rotación.
Cada trimestre, las viviendas de las secciones de un determinado turno de rotación son renovadas en su totalidad.

Diseño muestral de encuestas demográficas

Estimadores

El proceso habitual para la obtención de estimadores es:

1. *Estimador insesgado de expansión (Horvitz-Thompson)*: Compensa las desiguales probabilidades de selección
2. *Corrección de la falta de respuesta*: Corrige el sesgo producido en las estimaciones por la falta de respuesta (total)
3. *Calibrado con fuentes externas*: Reduce la varianza de las estimaciones mediante la utilización de fuentes externas

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

Diseño muestral de encuestas demográficas

1.- **Estimador insesgado de expansión(H-T)**

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

$$P(V_{ijh}) = P(S_{jh}).P(V_{ijh}/S_{jh}) = K_h.\frac{V_{jh}}{V_h}.\frac{m}{V_{jh}} = \frac{K_h.m}{V_h} = \frac{v_h^t}{V_h}$$

Siendo v^t_h el número *teórico* de viviendas de la muestra en el estrato *h*

Por tanto el estimador H-T tendrá la expresión:

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

Diseño muestral de encuestas demográficas

2.- Corrección de la falta de respuesta

La probabilidad de respuesta por estrato la podemos estimar por:

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

Donde v_h representa la muestra efectiva de viviendas en el estrato h

Por tanto el estimador corregido será:

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

Diseño muestral de encuestas demográficas

3.- Calibrado con fuentes externas(1)

Se utiliza, en primer lugar, un *estimador de razón separado* que toma como variable auxiliar las *proyecciones de población de 16 o más años*, a mitad del trimestre(P_h)

$$\hat{Y}_{Cal1} = \sum_h \frac{\hat{Y}_{H-TCorr(h)}}{\hat{P}_{H-TCorr(h)}} \cdot P_h$$

Es decir:

$$\hat{Y}_{Cal1} = \sum_h \frac{\frac{V_h}{v_h}\sum_{i\in h} y_i}{\frac{V_h}{v_h}\sum_{i\in h} p_i} \cdot P_h = \sum_h \frac{P_h}{p_h} \cdot y_h = \sum_S d_k \cdot y_k$$

Diseño muestral de encuestas demográficas

3.- Calibrado con fuentes externas(2)

En el segundo ajuste con fuentes externas se utilizan habitualmente las siguientes variables referidas a la población de 16 o más años en cada comunidad autónoma:

- *Proyecciones de población por grupos de edad y sexo(22) quinquenales*
- *Proyecciones de población por nacionalidad(españoles y extranjeros)*
- *Totales de población por provincia*
- *Número de hogares por tamaño*
- ……

Diseño muestral de encuestas demográficas

3.- Calibrado con fuentes externas(2)

Llamando x_j a cada una de las 'p' variables auxiliares (j=1,...,p), y X_j al total conocido en la comunidad autónoma, lo normal es que la muestra no sea equilibrada:

$$\hat{X}_j \neq \sum_{k \in s} d_k \, x_{jk}$$

Objetivo del calibrado: Obtener unos nuevos pesos w_k, lo más parecidos posible a los pesos d_k , que equilibren la muestra, es decir que verifiquen:

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

Diseño muestral de encuestas demográficas

3.- Calibrado con fuentes externas(2)

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

CALMAR es una macro pública de SAS
Disponible en: www.insee.fr
Nomenclatures, Definitions, Méthodes
Outils Statistiques

Desarrollo informático: Olivier Sautory(INSEE)
Teoría: Särndal, Deville y Sautory("Generalized Raking Procedures in Survey Sampling" JASA 1993 Vol.88, No423)

INSTITUTO NACIONAL DE ESTADISTICA

E.P.A. Segundo trimestre 2001

Resultados nacionales

	Datos	Impacto de los efectos (acumulativo por columnas)			
	publicados	Nuevas	Reponderación	Def. regulación	Diferencia
	(A)	poblaciones		1897/2000 (B)	(B-A)
1. Población de 16 años y más por sexo y relación con la actividad económica			(Valores absolutos en miles)		
AMBOS SEXOS					
Población de 16 años y más	32.926,8	33.651,6	33.651,5	33.651,5	724,7
Activos	16.898,7	17.283,2	18.214,5	17.709,9	811,2
- Ocupados	14.706,6	15.051,5	15.876,6	15.876,6	1.170,0
- Parados	2.192,1	2.231,7	2.337,9	1.833,3	-358,8
Inactivos	15.982,6	16.321,9	15.390,5	15.895,1	-87,5
Tasa de actividad	51,32	51,36	54,13	52,63	1,3
Tasa de paro	12,97	12,91	12,84	10,35	-2,6
VARONES					
Población de 16 años y más	15.844,7	16.192,5	16.344,9	16.344,9	500,2
Activos	10.113,5	10.341,4	10.977,4	10.794,9	681,4
- Ocupados	9.204,2	9.415,5	10.006,7	10.006,7	802,5
- Parados	909,3	925,9	970,7	788,2	-121,1
Inactivos	5.685,7	5.804,6	5.321,0	5.503,5	-182,2
Tasa de actividad	63,83	63,87	67,16	66,04	2,2
Tasa de paro	8,99	8,95	8,84	7,30	-1,7
MUJERES					
Población de 16 años y más	17.082,1	17.459,1	17.306,7	17.306,7	224,6
Activas	6.785,2	6.941,8	7.237,2	6.915,0	129,8
- Ocupadas	5.502,4	5.636,0	5.869,9	5.869,9	367,5
- Paradas	1.282,8	1.305,8	1.367,2	1.045,1	-237,7
Inactivas	10.296,9	10.517,2	10.069,5	10.391,6	94,7
Tasa de actividad	39,72	39,76	41,82	39,96	0,2
Tasa de paro	18,91	18,81	18,89	15,11	-3,8

INSTITUTO NACIONAL DE ESTADISTICA

Encuestas demográficas	Encuestas a Empresas
• Marco de areas y de lista	• Marco de lista
• Problemas de cobertura del marco	• Problemas del marco: unidades mal clasificadas
• Muestreo multietápico	• Muestreo monoetápico
• Afijación de compromiso	• Afijación óptima
• Errores de muestreo: Métodos Indirectos	• Errores de muestreo: Métodos directos
• Recogida de datos por entr. personal y telefónica	• Recogida de datos por correo y apoyo telefónico y web
• Coste elevado	• Menor coste
• Variables cualitativas	• Variables cuantitativas
• Carga estadística menor	• Carga estadística mayor

Evaluación de la calidad de los datos

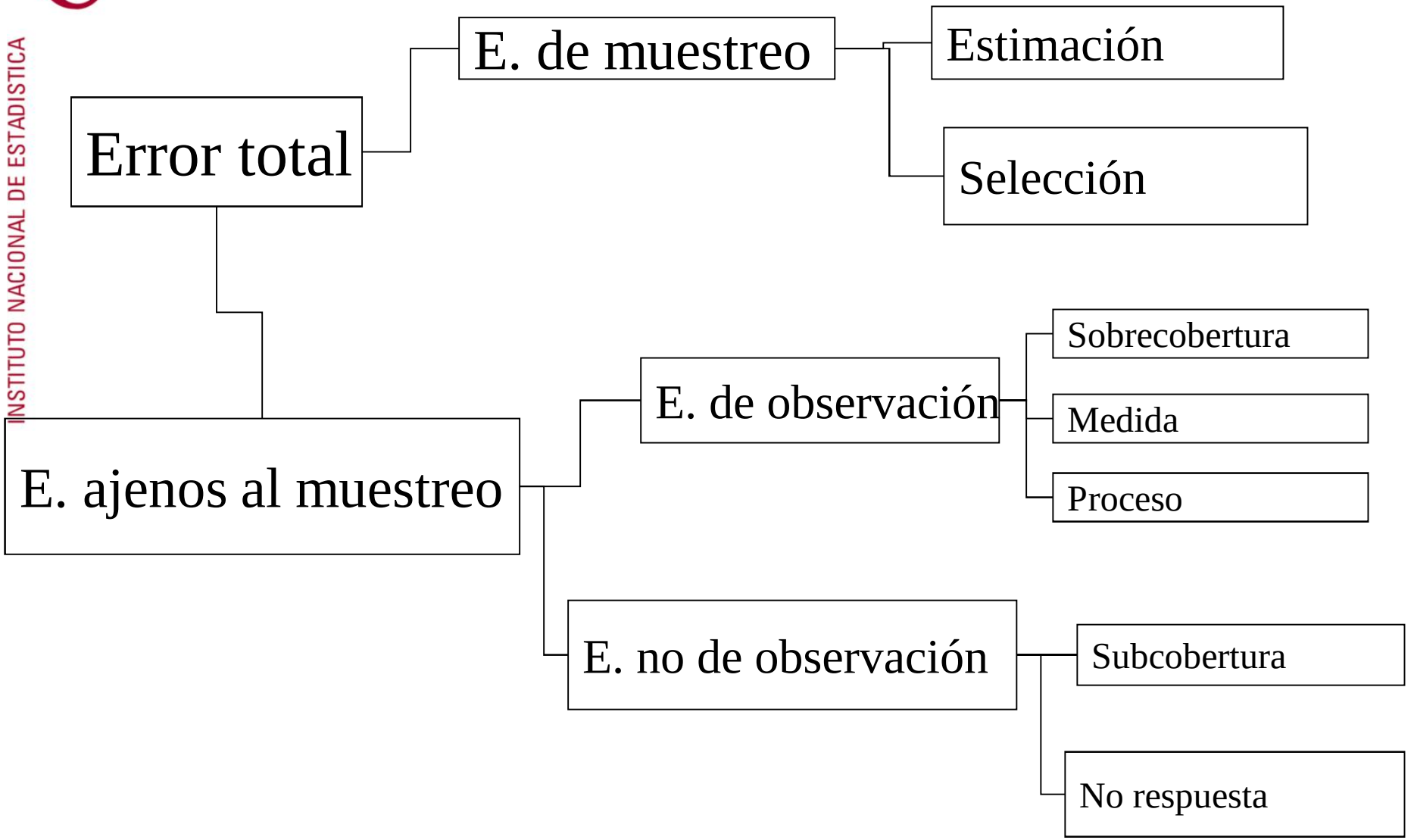
Errores de muestreo: Se producen por el hecho de observar una parte de la población, la muestra. Si el muestreo es probabilístico se pueden medir. Disminuyen con el incremento del tamaño muestral.

Errores ajenos al muestreo: Producen sesgos en las estimaciones. Difíciles de medir. Frecuentemente *¡aumentan!* con la ampliación del tamaño muestral.

- **De observación:**
  - Sobrecobertura
  - De medida
  - De proceso
- **No de observación:**
  - Subcobertura
  - Falta de respuesta

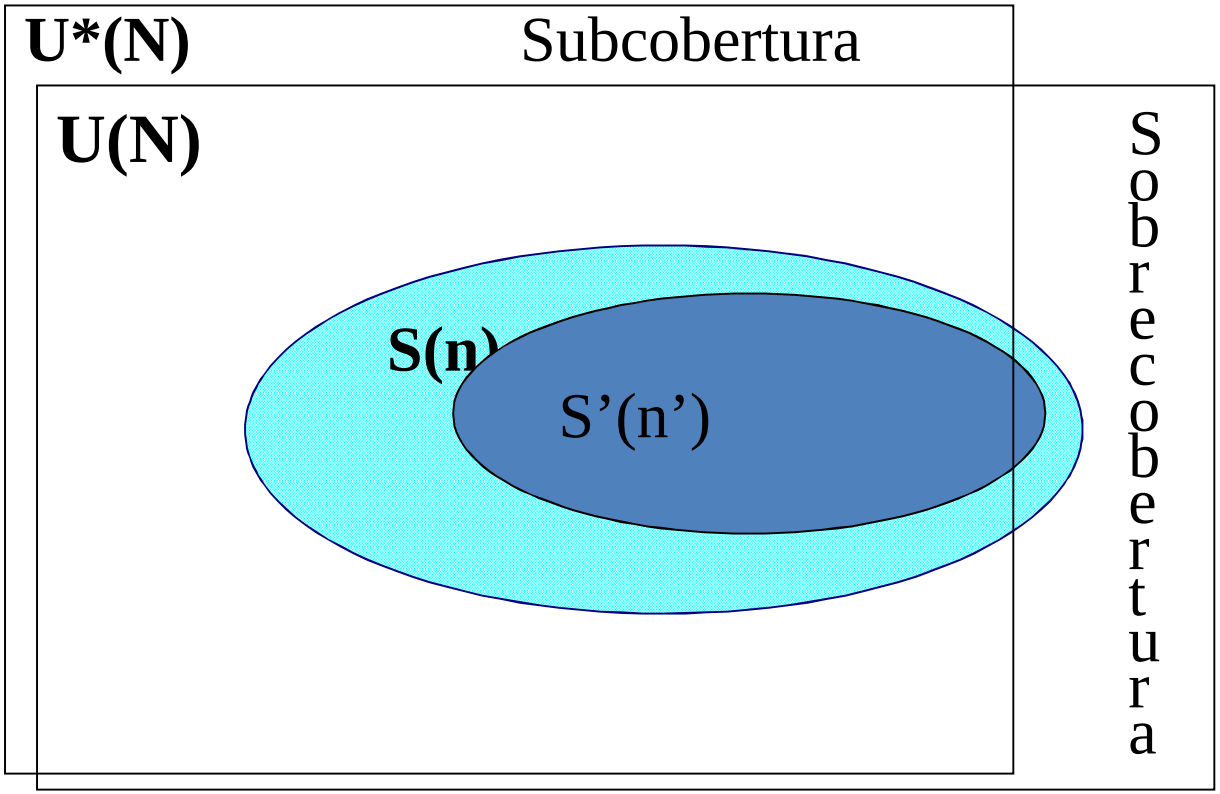
Evaluación de la calidad de los datos

Tipos de error en una encuesta



Evaluación de la calidad de los datos

Tipos de error en una encuesta



U: Marco

S: Muestra teórica

S': Muestra efectiva

U*: Población objetivo

Evaluación de la calidad de los datos

Errores de muestreo

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

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

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

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

Evaluación de la calidad de los datos

El conocimiento de la varianza de un estimador permite:

Al **diseñador** tomar decisiones entre diseños alternativos

Al **usuario** conocer el grado de fiabilidad de los datos

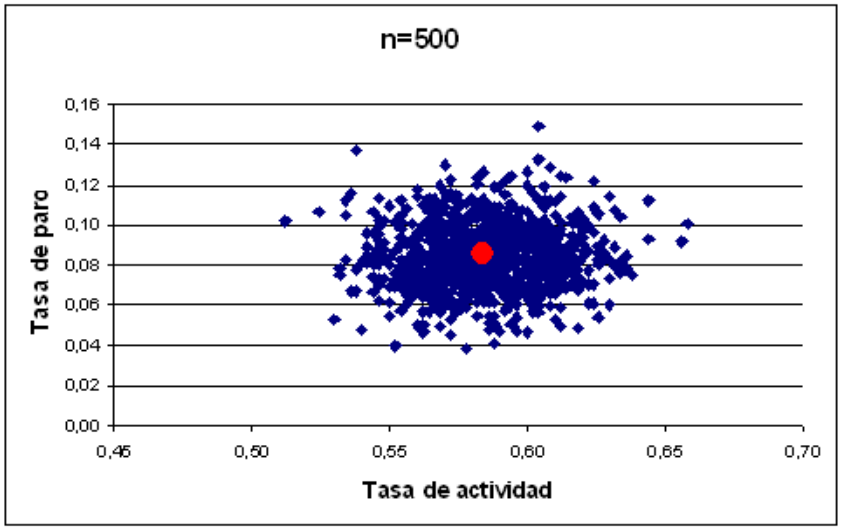
El error de muestreo proporciona al usuario un intervalo numérico que presenta una cierta confianza, medida en términos de probabilidad, de contener el valor verdadero que se desea estimar

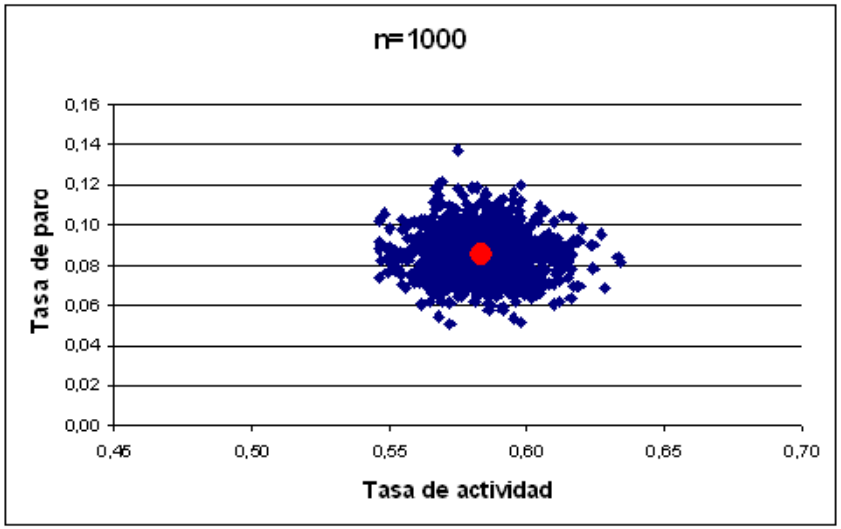
$$P\left\{ \hat{X} - 1.96\sigma(\hat{X}) < X < \hat{X} - 1.96\sigma(\hat{X}) \right\} = 0.95$$

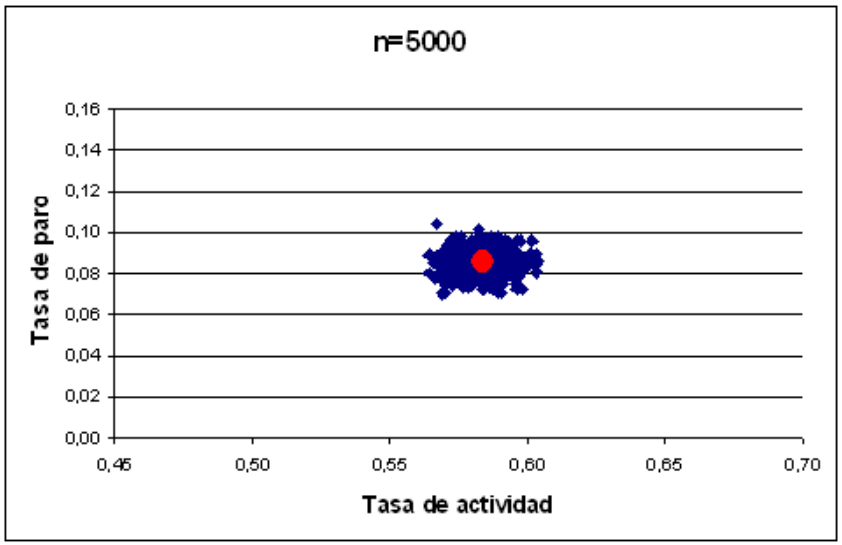
De cada 100 muestras obtenidas bajo el mismo diseño y condiciones generales, en 95 de ellas el intervalo de confianza obtenido contendría el valor verdadero (poblacional) con una probabilidad de 0.95

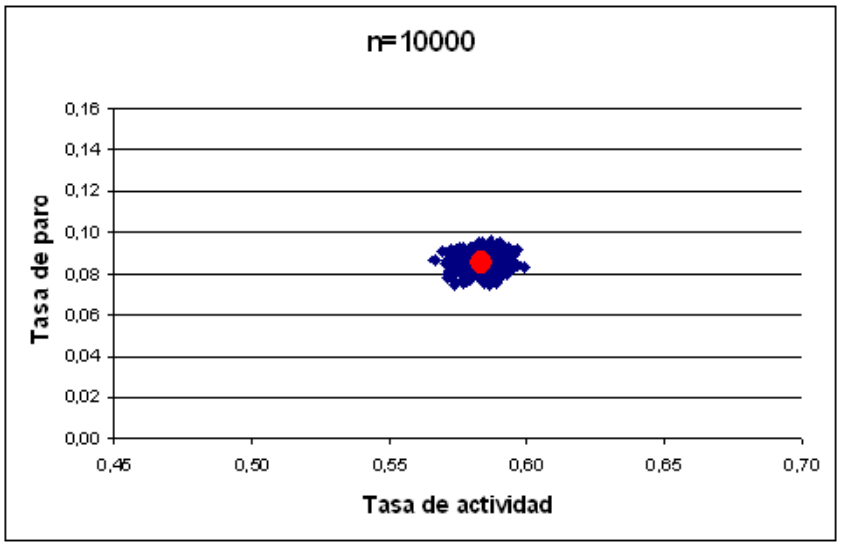
Lorca 2001.Simulación (m.a.s. 1000 iteraciones)

INSTITUTO NACIONAL DE ESTADISTICA









Evaluación de la calidad de los datos

INSTITUTO NACIONAL DE ESTADISTICA

Cálculo del error de muestreo

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

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

Evaluación de la calidad de los datos

Métodos Indirectos: Son aproximadamente insesgados y se basan en la formación de submuestras a partir de la muestra general.

La diferente forma de obtención de las submuestras da lugar a los diferentes métodos.

Los más utilizados son:

- Método de los grupos aleatorios
- " de los conglomerados últimos.
- " de las semimuestras reiteradas.
- " Jackknife
- " Bootstrap

Evaluación de la calidad de los datos

Errores de muestreo en la EPA

Se aplica el método de las semimuestras reiteradas

Consiste en:

•Obtención de sucesivas semimuestras de la muestra total.

•Estimación de la característica con cada semimuestr

El estimador de la varianza viene dado por:

$$\hat{V}\left(\hat{X}\right) = \frac{1}{r}\sum_{i=1}^{r}\left(\hat{X}_i - \hat{X}\right)^2$$ donde:

r es el número de semimuestras

$\hat{X}_i$ es la estimación con la i-ésima reiteración.

$\hat{X}$ es la estimación obtenida con la muestra completa.

Evaluación de la calidad de los datos

EPA

Errores de muestreo relativos,en porcentaje, de la población de 16 y más años según su relación con la actividad económica, por comunidades autónomas. Cuarto trimestre 2010

Comunidades autónomas	Activos	Ocupados	Parados			Inactivos
			Total	Buscan 1er empleo	Han trabaj. anterior.	
TOTAL	0,21	0,34	1,07	3,08	1,13	0,32
Andalucía	0,48	1,06	2,30	6,74	2,48	0,69
Aragón	0,66	1,89	8,74	28,68	8,42	0,91
Asturias (Principado de)	1,37	1,71	6,19	23,16	6,51	1,45
Balears (Illes)	1,23	2,13	4,53	22,80	4,60	2,12
Canarias	0,94	1,82	4,97	15,07	4,77	1,55
Cantabria	1,28	1,53	7,87	25,12	7,67	1,61
Castilla y León	0,57	0,98	3,86	10,93	4,10	0,70
Castilla-La Mancha	0,73	1,41	3,65	11,46	3,91	1,01
Cataluña	0,57	1,09	3,83	11,39	4,00	0,96
Comun. Valenciana	0,73	1,09	3,29	15,21	3,45	1,11
Extremadura	0,94	1,66	5,06	14,36	6,03	1,14
Galicia	0,51	0,87	3,56	10,30	3,41	0,61
Madrid (Comun. de)	0,77	1,23	4,82	14,29	4,98	1,45
Murcia (Región de)	0,89	1,47	3,88	17,26	4,17	1,51
Navarra (Com.Foral de)	0,92	1,12	8,33	42,73	9,03	1,37
País Vasco	0,76	1,07	5,45	17,18	5,89	1,04
Rioja (La)	1,43	1,46	7,65	20,41	7,89	2,09
Ceuta	7,28	11,34	17,12	32,10	15,84	8,55
Melilla	4,88	10,09	23,57	24,23	25,74	5,55

Evaluación de la calidad de los datos

INSTITUTO NACIONAL DE ESTADISTICA

Errores ajenos al muestreo

Son comunes a toda investigación estadística y se presentan en las distintas etapas de un proyecto estadístico:

-Antes de la toma de datos: Deficiencias del marco, insuficiencias en las definiciones o cuestionarios.

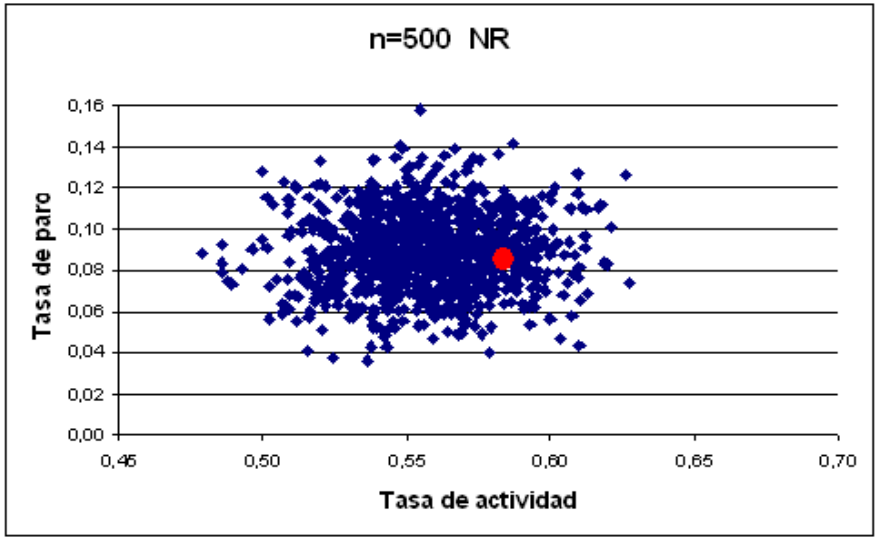
-Durante la recogida de la información: Defectos en la labor de los encuestadores, falta de respuesta por parte de los informantes, etc.

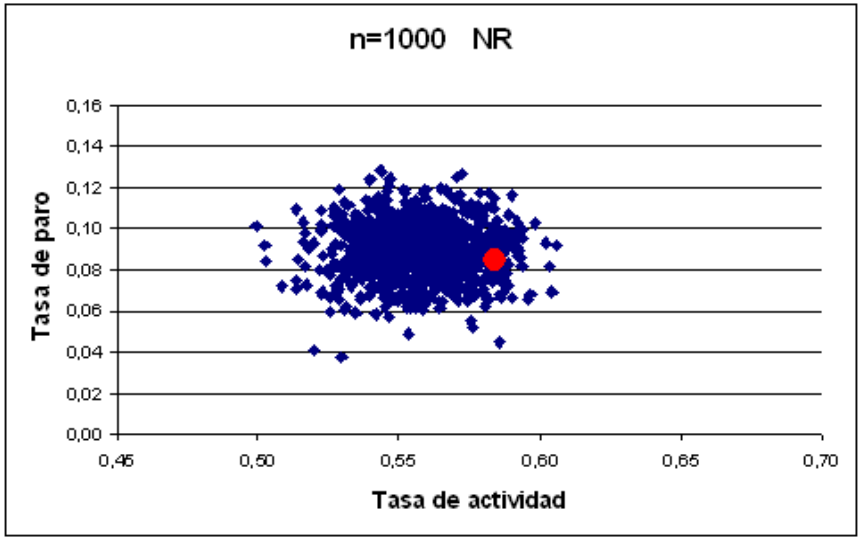
-Operaciones posteriores al trabajo de campo: Codificación, errores de grabación, etc.

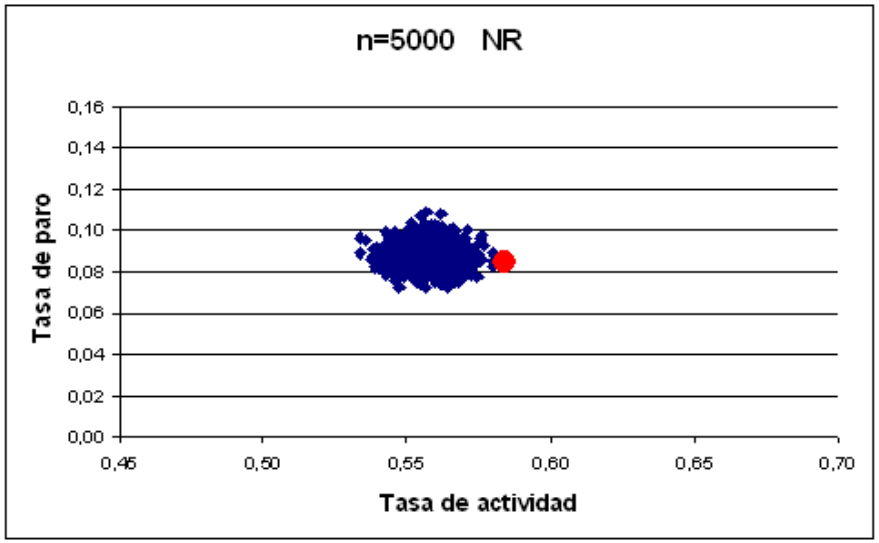
Producen sesgos difíciles de cuantificar.

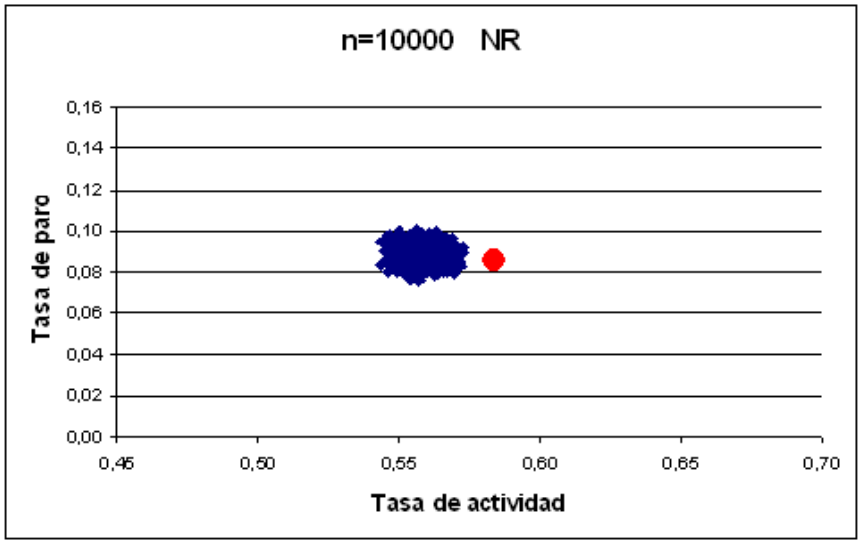
Lorca 2011.Simulación(m.a.s. 1000 iteraciones) No respuesta: PR(Varones)=0,7

INSTITUTO NACIONAL DE ESTADISTICA









Evaluación de la calidad de los datos

Errores ajenos al muestreo

EDAD07. Distribución de las viviendas *titulares* encuestadas y con incidencias por comunidades autónomas

Comunidades autónoma	Viviendas titulares												
	Total	Encuestadas		Con incidencia									
				Total		Falta de resp.		No encuestab.		Inaccesible		Selec. ant.	
	Nº	Nº	%	Nº	%	Nº	%	Nº	%	Nº	%	Nº	%
Total	96.412	63.813	66,2	32.599	33,8	21.023	21,8	11.238	11,7	97	0,1	241	0,2
Andalucía	20.040	13.526	67,5	6.514	32,5	4.050	20,2	2.397	12,0	32	0,2	35	0,2
Aragón	4.906	3.212	65,5	1.694	34,5	946	19,3	731	14,9	5	0,1	12	0,2
Asturias (Principado de)	1.928	1.416	73,4	512	26,6	354	18,4	158	8,2	0	0,0	0	0,0
Balears (Illes)	1.930	1.206	62,5	724	37,5	448	23,2	272	14,1	1	0,1	3	0,2
Canarias	3.862	2.511	65,0	1.351	35,0	835	21,6	505	13,1	4	0,1	7	0,2
Cantabria	1.575	969	61,5	606	38,5	408	25,9	197	12,5	0	0,0	1	0,1
Castilla y León	11.424	7.466	65,4	3.958	34,6	2.207	19,3	1.674	14,7	3	0,0	74	0,6
Castilla-La Mancha	6.493	4.272	65,8	2.221	34,2	1.203	18,5	997	15,4	7	0,1	14	0,2
Cataluña	9.818	6.490	66,1	3.328	33,9	2.544	25,9	746	7,6	11	0,1	27	0,3
Comunidad Valenciana	7.028	4.478	63,7	2.550	36,3	1.668	23,7	874	12,4	2	0,0	6	0,1
Extremadura	2.804	1.877	66,9	927	33,1	504	18,0	412	14,7	0	0,0	11	0,4
Galicia	6.142	4.275	69,6	1.867	30,4	1.064	17,3	792	12,9	3	0,0	8	0,1
Madrid (Comunidad de)	6.156	3.544	57,6	2.612	42,4	2.303	37,4	293	4,8	16	0,3	0	0,0
Murcia (Región de)	2.460	1.654	67,2	806	32,8	522	21,2	277	11,3	0	0,0	7	0,3
Navarra (Comunidad Fo	2.817	2.322	82,4	495	17,6	182	6,5	303	10,8	1	0,0	9	0,3
País Vasco	4.573	2.944	64,4	1.629	35,6	1.318	28,8	282	6,2	12	0,3	17	0,4
Rioja (La)	1.402	1.004	71,6	398	28,4	234	16,7	158	11,3	0	0,0	6	0,4
Ceuta y Melilla	1.054	647	61,4	407	38,6	233	22,1	170	16,1	0	0,0	4	0,4