Contribución al equilibrio de varias características deseables (*)

por F. AZORIN y J. L. SANCHEZ-CRESPO
Instituto Nacional de Estadística, España

1. CARACTERISTICAS DESEABLES (DESIDERATA) EN EL CAMPO DE LA ESTADISTICA

Los resultados estadísticos se consideran aceptables en tanto cumplan algunos requisitos básicos. Entre éstos deben destacar los siguientes:

a) *Acuracidad:* Proximidad a un valor verdadero. La acuracidad de un estimador incluye precisión e insesgamiento.

b) *Pertinencia:* Capacidad de los resultados estadísticos para completar ciertas lagunas en el conocimiento de un fenómeno especial.

c) *Oportunidad:* Utilidad de un resultado estadístico en función de su disponibilidad en el tiempo. Incluye: *puntualidad* (fijando límite de tiempo), *rapidez* (brevedad del tiempo transcurrido entre la primera etapa de producción de datos y su aprovechamiento) y *actualidad* (tiempo transcurrido desde su referencia).

d) *Accesibilidad:* Aptitud para ser conseguidos, recuperados, admitidos y comprendidos.

e) *Detalle y cobertura:* Aptitud para proporcionar información particularizada y circunstancias relacionadas, en profundidad y extensión.

f) *Integración:* Comparabilidad y otros atributos sistémicos.

g) *Economía:* Absoluta o relativa (eficiencia).

(*) Artículo presentado a la sesión plenaria de la Conferencia de Estadísticos Europeos, Ginebra, junio (1980), con el título «Contribution to the balancing of various desiderata attributes». Traducido por M.ª Rosa Villagrasa Novoa.

Estos requisitos generalmente se contraponen en el sentido de ser difícil o imposible mejorar uno de ellos sin detrimento de los demás. Puede llegarse a una situación de equilibrio, como resultante de la consideración simultánea de todos ellos, o bien pueden establecerse restricciones con valores mínimos de uno o varios de dichos requisitos.

En particular, una restricción muy importante en la práctica es el presupuesto disponible.

Asimismo, creemos que en la investigación debería tenerse en cuenta la *inocuidad*, en el sentido de protección del derecho a la intimidad y limitación de carga a los respondientes.

Finalmente, conviene mencionar la *credibilidad* de los resultados, como consecuencia de la imagen que tenga en los usuarios, el organismo productor de estadísticas.

Pasaremos ahora revista a estos requisitos con relación a la Encuesta española de la Población Activa.

Inocuidad: La carga a los respondientes, consistente en seis entrevistas trimestrales a lo largo de año y medio, no parece excesiva. No obstante, hay que tener en cuenta que en el programa de evaluación de la calidad de los datos se realizan una o dos entrevistas suplementarias en una submuestra. En cuanto al derecho de intimidad, éste queda protegido por la legislación española vigente.

Acuracidad: Se calculan y publican tanto los errores de muestreo como algunos de los ajenos al mismo. Como ejemplo, mencionaremos que la característica de desempleo, a nivel nacional, fue estimada en el 9,3 por 100 de la población activa (13.105.000 personas) para el último trimestre de 1979. A nivel provincial oscilan entre el 1,2 por 100 en Lérida y el 20,0 por 100 en Cádiz. Los errores relativos del muestreo fueron 1,9 por 100 para nivel nacional y 32,8 y 7,4 por 100, respectivamente, para las provincias mencionadas. Como indicador del sesgo de respuesta se calcula, en la encuesta de evaluación, el índice de cambio neto, siendo su valor −7,0. Finalmente, el número de personas idénticamente clasificadas en la entrevista original y en la repetida fue 64,6 por 100.

En cuanto a la cobertura, se obtuvo información directa en el 89,7 por 100 de las viviendas e información indirecta en 0,7 por 100 para negativas y 9,1 por 100 para ausentes.

El I. N. E. está realizando un estudio sobre acuracidad en relación al coste y oportunidad, y se está trabajando en la obtención de la encuesta piloto mencionada anteriormente.

Para los restantes requisitos, con excepción de la oportunidad, no parece que puedan alterarse sus valores con las restricciones presupuestarias actuales.

Oportunidad: En el estudio mencionado anteriormente se está tratando por un lado de reducir los intervalos de tiempo entre las distintas fases de producción y por otro se buscan mecanismos que permitan conocer los principales datos en un avance basado en una submuestra e incluso algún tipo de predicción.

2. CONSIDERACION DE LOS COSTES PARA LA ENCUESTA DE POBLACION ACTIVA

En relación con el estudio piloto mencionado en la sección 1, fueron considerados los siguientes datos básicos.

a) *Coste en porcentaje por fases*

Planificación (*)	5,38
Toma de datos	90,93
Proceso de datos	3,48
Publicaciones	0,21
	100,00

Esta tabla ha sido elaborada a partir de los ingresos brutos, por todos los conceptos, de las personas que a distintos niveles trabajan en la encuesta. Incluye como factor de carácter subjetivo el porcentaje de tiempo dedicado a la encuesta por ciertas personas, según la opinión de los jefes de unidad.

b) *Tiempo, en horas, por sección y vivienda*

	Horas de entrevista y localización por sección				Actualización por sección
	Entrevista	Localización	Total	Horas/viv.	Horas/sec.
Capital	7	3	10	$t_e = 0{,}5$	$t_a = 6{,}6$
Resto de la provincia	7	11	18	$t_e = 0{,}9$	$t_a = 6{,}6$

(*) Organización y mantenimiento del trabajo del campo no respuesta y procedimientos de imputación, retroalimentación («feed-back») para el diseño de la muesta, actualización del marco y probabilidad, etc.

c) *Costes medios por hora, en pesetas (*)*

	Agente	Jefe grupo
Capital	362	385
Resto de la provincia	385	407
Oficina	362	385

d) *Otros datos*

— Distancia media oficina-sección-oficina: D = 230,3 Km.

— Tiempo medio para recorrer un Km: t_k = 0,0125.

— Tiempo medio de proceso por cuestionario en la oficina provincial: t_d = 0,3 horas.

— Tiempo medio de actualización por sección: 6,6 horas.

— Coste por Km: C_k = 9 ptas.

— Coste medio de proceso por un cuestionario en la oficina provincial: C_d = 108,6 ptas.

Por cada tres agentes existe un jefe de grupo.

El número de agentes para esta encuesta es aproximadamente de 250.

A partir de estos datos hemos formado la siguiente tabla:

Tipo de operación	Entre secciones	En las secciones	En las viviendas
Entrevista, localización y actualización		$C_e \cdot t_a \cdot n$	$C_e \cdot t_e \cdot n\, \overline{m}$
Viaje	$C_e \cdot t_k \cdot D \cdot n$ $C_k \cdot D \cdot n$		
Depuración y otros trabajos de oficina			$C_d \cdot t_d \cdot n\, \overline{m}$

De la que se obtiene la siguiente función de costo:

$$C = \left\{\left(\frac{4}{3}C_e t_k + C_k\right) \cdot D + \frac{4}{3}C_e t_a\right\} \cdot n + \frac{4}{3}\left\{C_e t_e + C_d t_d\right\} n\, m$$

(*) 1 US $ = 70 ptas.

en la que utilizaremos el factor 4/3 para incluir el costo de los jefes de grupo.

La función de coste para el total provincial, excluida la capital, era:

$$C(P) = 10.164\ n + 1.412{,}4\ n\ \overline{m}$$

Análogamente se obtuvo la siguiente función de coste para la capital:

$$C(C) = 3.185{,}6\ n + 284{,}77\ n\ \overline{m}$$

donde n es el número de secciones y $n\ \overline{m}$ el número de viviendas.

El coste total de la recogida de datos es, por lo tanto:

$$C = C(P) + C(C)$$

3. EQUILIBRIO ENTRE COSTES Y ACURACIDAD

Estudiaremos por separado el conjunto de las provincias (excluyendo las capitales) y el de las 50 capitales.

Los pares de valores $(\overline{m},\ n)$ que satisfacen la función de coste por provincias en la proximidad del par (20; 1.000) utilizado actualmente por $\delta = 0{,}05$, $P = 0{,}38$ (*), son:

$\overline{m}$	n	$n\ \overline{m}$	$C^2(\hat{P}) \cdot 10^6$	$C(\hat{P}) \cdot 10^3$
35	645	22.575	195	14,0
30	731	21.930	182	13,5
25	845	21.125	170	13,0
20	1.000	20.000	159	12,6
15	1.225	18.375	151	12,3
10	1.582	15.820	149	12,2
12	1.417	17.004	148	12,1
5	2.230	11.150	175	13,2
2	2.957	5.914	289	17,0

donde $C^2(\hat{P})$ es la varianza relativa a cuadrado del coeficiente de variación. Vemos que el par (12; 1.417) corresponde a la mínima varianza relativa.

En el conjunto de las capitales de función de coste es:

$$C(C) = 3.185{,}6\ n + 284{,}77 \cdot n \cdot \overline{m}$$

(*) Valores utilizados en la función de costo de 1971.

y los pares ($\overline{m}$, n) que la satisfacen y las correspondientes varianzas relativas para análogos valores de δ y P son:

$\overline{m}$	n	$n\,\overline{m}$	$C^2(\hat{P}) \cdot 10^6$	$C(\hat{P}) \cdot 10^3$
35	1.350	47.250	93	9,6
30	1.514	45.420	88	9,3
25	1.724	43.100	83	9,1
20	2.000	40.000	79	8,9
15	2.382	35.730	78	8,8
10	2.944	29.440	80	8,9
5	3.854	19.270	102	10,1
2	4.730	9.460	181	13,5

En este caso, el par óptimo consiste en $\overline{m} = 15$ viviendas y $n = 2.382$ secciones.

Puede observarse que en ambos casos más que de valores óptimos podemos hablar de regiones óptimas, $5 < \overline{m} < 20$, $1,00 < n < 2.230$ en provincias; $5 < \overline{m} < 25$, $1.724 < n < 3.854$ en capitales. Parece por consiguiente útil para tomar una decisión considerar otros aspectos operativos que no son fáciles de incluir en un tratamiento matemático.

A continuación obtendremos los valores óptimos de m con la fórmula

$$\overline{m}_{op} = \sqrt{\frac{C_1}{C_2} \cdot \frac{1 - \delta}{\delta}}$$

para varios valores de δ

	m_{op}		
	$\delta = 0,01$	$\delta = 0,05$	$\delta = 0,09$
Provincias (excepto capitales)	27	11,7	8,5
Capitales	33	14,4	10,5

calculándose los correspondientes valores de n a partir de las funciones de costo.

Esta tabla pone de manifiesto el efecto, sobre el tamaño de la muestra, de consideraciones multidimensionales del problema.

RESUMEN

Se definen las siguientes características deseables (desiderata) para los resultados estadísticos: Acuracidad, Pertinencia, Oportunidad, Accesibilidad, Detalle y Cobertura,

Integración, Economía, y también Inocuidad y Credibilidad. Se consideran estos requisitos en relación con la Encuesta de Población Activa.

Palabras clave: Características deseables, desiderata, resultados estadísticos, definiciones, acuracidad, función de coste.

SUMMARY

The following desiderata of statistical results are defined: Accuracy, Relevancy, Timeliness, Accesibility, Detail and Coverage, Integration, Economy, and also Innocuousness and Credibility. These requirements are reviewed in relation to the Spanish Labor Force Survey.

Key words: Desiderata, statistical results, definitions, accuracy, cost functions.

AMS Subject classification: 62D05