

Decisión sobre medidas poblacionales desconocidas antes del muestreo

Autor: Vicente Manzano Arrondo

Departamento de Psicología Experimental

Facultad de Psicología de la Universidad de Sevilla

Av. San Francisco Javier s/n, 41005 · Sevilla

Resumen

La información sobre la población de interés puede no estar disponible antes de realizar una investigación por muestreo. En tales casos, deben tomarse decisiones con respecto a algunas funciones poblacionales para, por ejemplo, escoger un tamaño de muestra concreto. En este trabajo, se sugieren algunas estrategias para obtener cantidades numéricas concretas con respecto a las funciones poblacionales más comunmente requeridas.

Palabras clave: muestreo, parámetros desconocidos.

Introducción

El objetivo de toda investigación por muestreo es obtener información sobre la población de la que se extrajo la muestra. Para ello, el primer requisito es obtener una muestra representativa en la que realizar los cálculos cuyos resultados servirán para inferir acerca de la población. Asimismo, para obtener la muestra, la primera decisión se refiere a ¿cuántos elementos deben seleccionarse para llegar al objetivo con la seguridad y precisión requeridos? (Feigl, 1978; Teijeiro, 1990; Sudman, 1983; Kalton, 1987; Fink y Kosecoff, 1989; Henry, 1990; Czaja y Blair, 1996). Por último, para responder a esta pregunta, lamentablemente, es necesario contar con alguna información sobre la población (McCall, 1982), circunstancia que se ha venido a denominar *la paradoja de Friedman* (Azorín y Sánchez Crespo, 1986). «Nos encontramos así con la gran paradoja de la estimación y que es, que para poder estimar algo correctamente se precisa tener ya muchos conocimientos sobre ello o, dicho de otra manera, cuanto más sepamos de un dato, mejor podremos estimarlo» Salgado (1990:349).

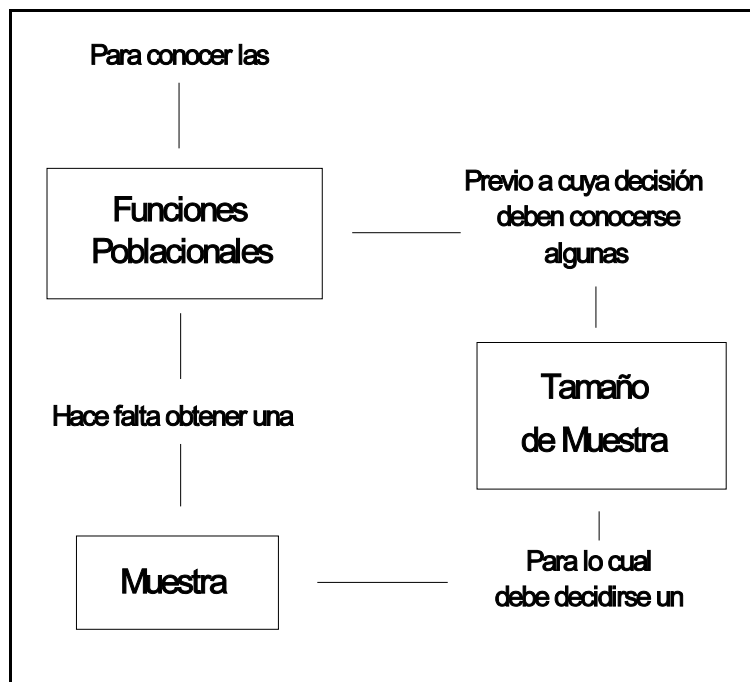


Figura 1: *Ciclo de decisión para el tamaño de la muestra, en función de algunas funciones poblacionales.*

El diagrama de la figura 1 ilustra esta problemática, mientras que el diagrama de la figura 2 presenta cómo se soluciona en la práctica. Lo cierto es que el tamaño de la muestra se acerca más a lo óptimo cuanto mayor es la información con que se cuenta de la población. Tal circunstancia implica que en los estadios iniciales de investigación en un campo de estudio, se exigirán muestras más grandes que las que se necesiten en momentos de desarrollo posterior, en los que el aumento del conocimiento sobre la población permitirá estimaciones precisas con muestras pequeñas.

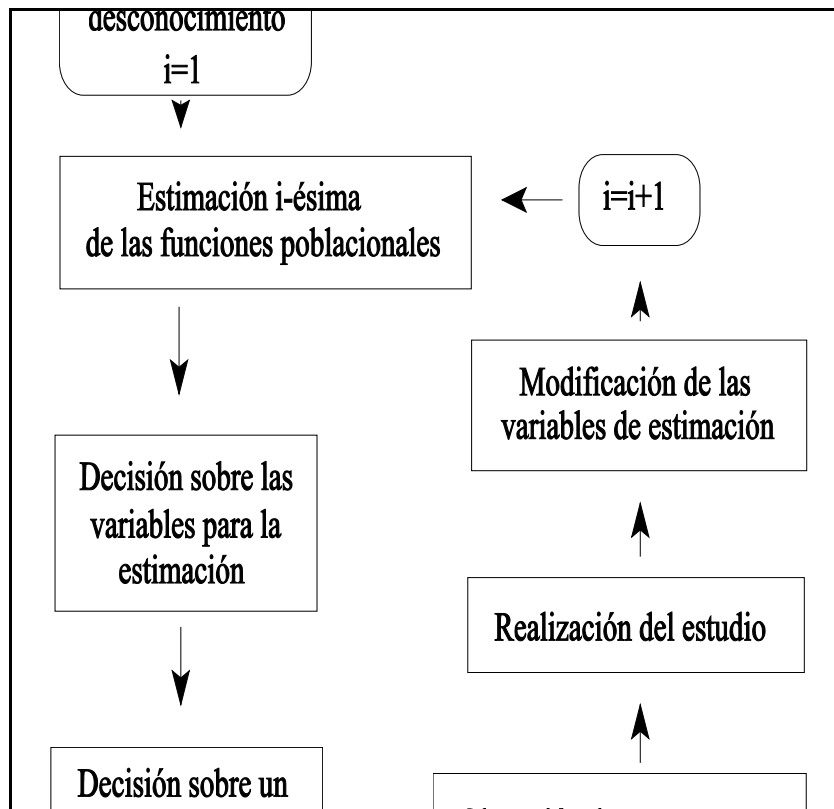


Figura 2: Ciclo de estimación de las funciones poblacionales a partir de los resultados obtenidos en sucesivos procedimientos de muestreo.

E 1

conocimiento sobre las funciones poblacionales previas están sujetas a incertidumbre (Kish, 1965; Wilburn, 1984), por lo que no puede aspirarse a conseguir un tamaño de muestra con

precisión absoluta, de tal forma que llega a afirmarse «decidir qué tamaño de muestra utilizar es casi siempre más una cuestión de juicio que de cálculo» (Hedges, 1980:61). Aún así, sí que podemos aspirar a realizar la mejor de las aproximaciones posibles, mediante la reducción de la incertidumbre, de tal forma que se minimice el monto de error global en el estudio.

En otros términos, se requieren herramientas que permitan la recogida de información útil para acotar los valores de las funciones poblacionales necesarias para el cálculo del tamaño de la muestra, puesto que «mientras más información se tenga inicialmente, acerca de una población, más fácil será proyectar una muestra que dé estimaciones exactas» (Cochran, 1976:29).

En este sentido cabría distinguir entre dos tipos de funciones poblacionales de interés en una investigación con muestreo:

1. *Medidas procedimentales*: son funciones poblacionales que interesan para tomar decisiones sobre las características de la muestra. En el contexto del presente estudio, las medidas procedimentales permiten obtener un tamaño de muestra.
2. *Medidas objetivo o meta*: son las funciones poblacionales que interesan estimar a partir de funciones calculadas sobre los datos de la muestra.

Así, en la estimación de la medida objetivo *media aritmética*, se requiere algún conocimiento previo de la medida procedimental *varianza*, sin la que no es posible decidir un tamaño óptimo de muestra. La figura 3 muestra la relación entre ambos tipos de funciones poblacionales.

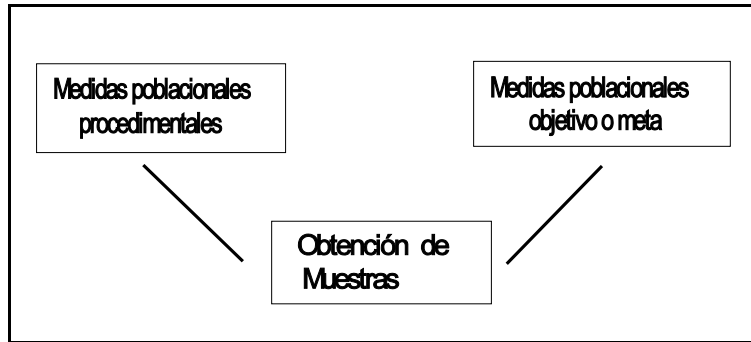


Figura 3: Flujo de información entre funciones poblacionales, según su interés.

Contamos con dos preguntas clave que responder para tomar decisiones sobre medidas poblacionales procedimentales. La primera de ellas es qué funciones poblacionales están implicadas en la decisión sobre el tamaño de la muestra. La segunda es cómo estimar sus valores. Los siguientes apartados intentarán responder ambas.

Medidas procedimentales implicadas

Las funciones poblacionales que sirven para decidir el tamaño de la muestra, están en función del contexto de inferencia y del modelo de selección de muestra a utilizar. La figura 4 muestra una propuesta de agrupación para las medidas procedimentales según las situaciones más frecuentes: modelos de muestreo aleatorio simple, estratificado, de conglomerados monoetápico y de conglomerados con submuestreo (bietápico). De igual forma, se han considerado tres objetivos de inferencia: problema univariable o de diferencia en grupos relacionados o independientes.

	Contexto de inferencia		
	Univariable	Grupos relacionados	Grupos independientes
Simple	σ, N	$\sigma_1, \sigma_2, \rho, N$	$\sigma_1, \sigma_2, w_1, w_2$
estratificado	σ_i, N_i	$\sigma_{1i}, \sigma_{2i}, \rho_i, N_i$	$\sigma_{1i}, \sigma_{2i}, w_{1i}, w_{2i}$
monoetápico	σ, δ, G, N_c	$\sigma_1, \sigma_2, \rho, \delta, G, N_c$	$\sigma_1, \sigma_2, \delta, w_1, w_2$
submuestreo	$\sigma, \delta, G, N_c, n_c$	$\sigma_1, \sigma_2, \rho, \delta, G, N_c, n_c$	$\sigma_1, \sigma_2, \delta, w_1, w_2, n_c$

Figura 4: Relación de medidas procedimentales según el contexto de inferencia y el modelo de extracción de muestras. La variable n_c no puede ser considerada una medida procedimental, pero constituye el factor diferenciador en cuanto a la existencia o no de submuestreo en los modelos de extracción por conglomerados.

Los símbolos utilizados corresponden a:

N	Número de unidades elementales
N_c	Número de conglomerados
n_c	Número de conglomerados (en la muestra)
σ	Varianza
σ_j	Varianza en el grupo j
ρ	Correlación entre grupos
w_i	Peso del grupo i
G	N / N_c
δ	correlación u homogeneidad entre conglomerados
X_i	Medida X en el estrato i

Siempre se requiere el conocimiento de, al menos, una varianza en la población, cuantía que casi siempre es desconocida (Gillett, 1989). Si, además, no se parte de un modelo para poblaciones infinitas, es necesario contar con el tamaño de la población. Éste es el modelo más simple, correspondiente a la estimación univariable en el muestreo aleatorio simple. Conforme se varía el contexto de inferencia o el modelo de muestreo, los requerimientos de información aumentan.

Trataremos las medidas procedimentales en tres categorías independientes: tamaños (N , G , w), varianzas (σ) y otros (ρ y δ). Con el mismo orden, serán consideradas en los siguientes apartados.

Estimación de tamaños

Es, sin duda, el apartado de más fácil solución. El número de unidades de la población es conocido. Cuando esta información no es precisa, ocurre que el contexto implica a poblaciones de gran tamaño, en cuyo caso errores absolutos de cuantía apreciable, no tienen consecuencias sensibles en el proceso de estimación.

El problema puede surgir en los modelos de muestreo con partición (Manzano, 1997a). Se requiere conocer el número de unidades en cada clase. Si éste es desconocido, no existen procedimientos deductivos para afrontar una solución, sino únicamente:

1. Suponer que todas las clases (sean estratos o conglomerados) cuentan con el mismo número de unidades.
2. Realizar un muestreo polietápico. En cada etapa, se consideran conglomerados a un nivel tal que permita conocer el número de unidades de cada uno. Únicamente dentro de los conglomerados seleccionados se realiza una investigación que lleve a la planificación

de la siguiente etapa donde, de nuevo, se considera una partición tal que permita conocer el tamaño de cada clase.

Cuando el muestreo es estratificado, el investigador puede carecer de la información necesaria para establecer la partición en cada conglomerado seleccionado. En tales casos cabe realizar una estimación del tamaño de clase en base a unidades de un nivel superior. Así, por ejemplo, si el investigador no cuenta con los tamaños por estrato en una partición bidimensional por edades y sexos en municipios, puede utilizar la tabla de contingencia a nivel provincial y establecer una equivalencia proporcional en cada clase.

En definitiva, o bien el investigador pone en marcha un procedimiento de recogida de datos, mediante un modelo polietápico, o bien aplica la suposición de igualdad de tamaños.

Varianza poblacional

Si bien la acotación de valores para las medidas procedimentales no es un tema que haya generado un número sensible de trabajos de investigación o reflexión, si es cierto que la mayoría de éstos se refieren a la suposición de valores para la varianza poblacional, omnipresente en cualquier modelo y contexto de inferencia cuando se requiere el cálculo del tamaño para la muestra.

Barnett (1974:33-35) y Cochran (1976:112-116) indican cuatro soluciones para la obtención de un valor para σ :

1. Realizar un estudio piloto. En éste se recoge información que si bien no goza de las máximas garantías (es una muestra muy pequeña donde, además, se ponen a prueba el cuestionario, los encuestadores, el procedimiento, etc...), es útil para acotar los valores razonables para la varianza en la población.
2. Obtener una *premuestra*. Se toma una muestra previa en la que se realizan las

mediciones con las mismas características que en el estudio posterior. A diferencia con el estudio piloto, la premuestra forma parte de la muestra final, participando de ésta. Los resultados obtenidos con estos primeros datos permiten realizar una estimación inicial para la varianza poblacional que aconseje la extracción de un número determinado de unidades, además de las ya obtenidas en la muestra previa. Así, si la estimación de la varianza en los primeros n_1 datos lleva a la decisión de un tamaño n_2 para la muestra, será necesario extraer $n_2 - n_1$ unidades más.

Otro formato de partición de una muestra, distinta al concepto mencionado de *pre muestra* es la que aconseja Kish (1965): en situaciones de incertidumbre lo más conveniente es seleccionar una muestra razonablemente grande, con base en las estimaciones más pesimistas para las varianzas implicadas. El tamaño se divide en dos: uno inicial y otro, denominado *suplemento*. Tras la recolección y análisis de datos de la muestra inicial, se decide si procede o no continuar con el suplemento planificado. Esta estrategia es preferible a interrumpir bruscamente el proceso de selección cuando se ha conseguido un n suficiente, ya que las unidades no seleccionadas introducen un sesgo importante.

3. Encontrar información sobre el tema en los resultados de investigaciones previas. Si han existido, en ellas se obtienen estimaciones útiles.

4. Basándose en algunas suposiciones sobre la distribución de la variable en la población, pueden acotarse los extremos posibles, admisibles o razonables para la varianza poblacional.

De estas cuatro categorías, podemos distinguir a su vez dos subgrupos. Las tres primeras requieren un proceso de investigación previo, bien sea recogiendo datos de la población directamente (soluciones 1 y 2) o indirectamente a través de trabajos anteriores (solución 3). La

cuarta categoría es diferente y centra nuestro interés en el presente trabajo. La suposición de algunas características definitorias de la distribución de la variable en la población, permite acotar valores para la varianza, de tal forma que se llegue a una solución satisfactoria sin necesidad de recurrir a extracciones previas de datos que encarecen el estudio y distorsionan los cálculos sobre costes, tiempo y tamaños de muestra. Aún así, se trata de la estrategia de solución más burda, puesto que su base está asentada en suposiciones, no en informaciones empíricas.

Conforme el investigador cuente con una mayor cantidad y calidad de información, la acotación será más precisa y ello redundará en un tamaño de muestra inferior con respecto a las situaciones de mayor incertidumbre. Por tanto, estableceremos las siguientes categorías:

1. Conocimiento de σ a partir de investigaciones previas realizadas específicamente (estudio piloto o premuestra) o independientemente (búsqueda bibliográfica).
2. Conocimiento sobre algunas características de la distribución de la variable en la población.
3. Conocimiento sobre un valor aproximado para la media poblacional y/o de los valores extremos.
4. Desconocimiento absoluto.

En el primer caso, la solución para el valor de σ pasa por traducir los valores obtenidos en el estudio piloto, premuestra o investigaciones previas, en función del modelo de muestreo utilizado, realizando una estimación puntual de la varianza. Los desarrollos para las otras tres situaciones son los que nos interesan en este apartado.

Los contenidos que siguen parten de la estimación de medias, es decir, del contexto en el que están involucradas variables continuas. Por separado y de forma previa, abordamos la estimación de proporciones, por ser de solución más sencilla y establecer conclusiones que serán utilizadas en el caso de la estimación de medias.

Estimación de proporciones

La situación de máxima incertidumbre queda definida como el desconocimiento total de la distribución de valores en la población, así como de un valor aproximado para π . En tal caso, aún es posible obtener alguna información útil, derivada de las propiedades algebraicas de las proporciones. Nuestra primera preocupación es acotar el valor de σ .

Derivando la función sigma cuadrado e igualando a 0, obtendremos el punto en que la función muestra pendiente nula y, por tanto, cuenta con su máximo (no un mínimo, pues es función enteramente positiva, con límites nulos):

$$D\sigma^2 = D(\pi^2 \&\pi) = 2\pi \&1 = 0 \quad \text{en} \quad \pi = .5 \quad \text{y} \quad \sigma = .5$$

Por tanto, $0 \leq \sigma \leq .5$

Así pues, en el caso de absoluto desconocimiento sobre cualquiera de los aspectos que definen la distribución poblacional de la variable de interés en la inferencia, puede establecerse que la desviación tipo $\sigma = 0.5$, o bien $\sigma^2 = 0.25$. Se trata de una salida conservadora por cuanto que considera el máximo valor posible para la varianza, es decir, el valor de ésta que llevará a la decisión de un tamaño de muestra máximo. Esta circunstancia motivará que el error de precisión decidido de antemano sea una cota superior.

No obstante, lo usual es contar con alguna información previa sobre el valor esperable de π , en el sentido de que pueden establecerse sus extremos esperables:

$$\min(\pi)=m \quad \text{y} \quad \max(\pi)=M$$

En tal caso, el máximo valor esperable para σ^2 se consigue a partir de la cuantía del intervalo (m, M) que más se aproxime a $\pi=0.5$. Es decir:

* Si $0.5 \in (m, M)$ $\sigma^2 = 0.25$

* Si $0.5 \notin (m, M)$

$$* \text{ Si } *m - 0.5* < *M - 0.5* \text{ Y } \sigma^2 = m(1-m)$$

(o bien, si $[m-0.5]>0$)

$$* \text{ Si } *m - 0.5* > *M - 0.5* \text{ Y } \sigma^2 = M(1-M)$$

(o bien, si $[M-0.5]<0$)

Conocimiento sobre algunas características de la distribución de la variable en la población

No es infrecuente la situación en la que el investigador trabaja con variables continuas de las que sospecha, con cierto fundamento, que siguen un determinado modelo de distribución en la población de interés. A su vez, los modelos asumidos, por lo general, no constituyen un conjunto amplio de alternativas. Lo más frecuente es considerar la distribución normal o alguna variación ligeramente asimétrica.

Por otro lado, no son todas las características de un modelo de distribución lo que nos interesa, sino únicamente aquellas que atañen a la varianza. Al respecto, consideraremos únicamente las cuatro siguientes alternativas:

1. Máxima varianza. La distribución de los datos entre los valores máximo y mínimo se establece saturando únicamente éstos, de tal forma que la mitad de los datos muestran el valor mínimo y la otra mitad el valor máximo. En tal caso, la distancia cuadrática a la media coincide en todos los casos, con lo que:

$$\sigma^2 = \frac{\sum (X_i - \bar{X})^2}{N} = \frac{1}{N} \left(2 \cdot \frac{N}{2} \left[\max - \frac{\max + \min}{2} \right]^2 \right) = \frac{(\max - \min)^2}{4}$$

2. Modelo de distribución uniforme. Representa a conjuntos con gran dispersión, puesto

que los datos no se agolpan en torno a ningún centro, sino que permanecen equitativamente dispersos en toda la amplitud de posibles valores para la variable. La varianza de la distribución uniforme queda definida (por ejemplo, Gutiérrez Cabría [1978:145]) por:

$$\sigma^2 = \frac{(\max \& \min)^2}{12}$$

3. Modelo de distribución normal. Posiblemente la tendencia más esperable en las variables continuas para multitud de características de interés. Con la misma amplitud esperable de valores, representa una varianza sensiblemente menor que la distribución uniforme. Suponiendo que los valores máximo y mínimo esperables para la variable constituyen la gran mayoría de las posibilidades (considerando que la amplitud de la normal es teóricamente infinita), puede establecerse una distancia estandarizada de valor $Z = 2.58$, correspondiente al 99% central. En tal caso:

$$Z_i = \frac{X_i - \mu}{\sigma} \quad \text{Y} \quad \sigma = \frac{X_i - \mu}{Z_i} = \frac{\frac{\max \& \min}{2}}{Z_i} = \frac{\max \& \min}{5.16} \quad \text{Y}$$

$$\sigma^2 = \frac{(\max \& \min)^2}{27}$$

4. Mínima varianza: En esta situación, todos los datos coinciden con la media, salvo dos valores extremos, definidos por el máximo y el mínimo. En tal caso, únicamente hay dos

puntuaciones diferenciales no nulas, con lo que:

$$\sigma^2 = \frac{\sum (X_i - \bar{X})^2}{N} = \frac{2}{N} \left(\max \& \frac{\max - \min}{2} \right)^2 = \frac{(\max - \min)^2}{2N}$$

De los modelos 1 a 4 se fluye de máxima a mínima varianza. No obstante, los contextos 1 y 4 son suficientemente extremos como para considerar que no ocurren en la práctica, con lo que son los modelos uniforme y normal los razonablemente aplicables.

Conocimiento sobre un valor aproximado para la media poblacional y/o de los valores extremos

Como continuación del subapartado anterior, si el investigador desconoce la distribución de base para la variable en la población, pero es capaz de establecer un intervalo de valores máximo y mínimo, puede optar por dos soluciones alternativas admisibles:

1. Aproximación intermedia: considerar una distribución mixta, a medio camino entre la uniforme y la normal con respecto al valor supuesto para la varianza. En tal caso:

$$\sigma^2 = \frac{(\max - \min)^2}{19.5}$$

2. Aproximación conservadora: considerar la distribución uniforme cuya estimación de la varianza es elevada. Una aproximación basada en la máxima varianza quizá resulte excesivamente conservadora, más aún considerando lo poco probable de la situación.

Una alternativa razonable es utilizar algún valor considerado para la media poblacional. Si el investigador es capaz de aventurar dos valores extremos para la amplitud de la variable en la población, una estrategia derivable es tomar el centro de esta amplitud como valor probable

para la media poblacional, es decir:

$$\tilde{\mu} = \frac{\max \& \min}{2}$$

¿Cómo operar con la estimación supuesta para μ ? Podemos recurrir a una analogía con el tratamiento de las proporciones, mediante una estrategia que compare el valor medio (π) con la varianza. Nuestra aproximación consiste en encontrar el valor que hace máximo el cociente entre la varianza y la proporción poblacional, de tal forma que escoja un valor conservador para σ^2 a partir de π . Es decir:

$$\max \left(\frac{\sigma^2}{\pi} \right) = \max \frac{\pi (1 \& \pi)}{\pi} = \max (1 \& \pi) = 1 \& \min (\pi) = 1$$

Por lo que, la máxima discrepancia entre la varianza y la estimación de la media poblacional, utilizando la analogía con la relación establecida para las proporciones, será:

$$\frac{\sigma^2}{\tilde{\mu}} = 1 \quad \text{Y} \quad \sigma^2 = \tilde{\mu}$$

Si el investigador cuenta con un valor acotado de μ , bastará con la consideración de su cota superior. Si, en lugar de ello, puede acotar los extremos de la distribución en la población, aunque desconozca otras características de ésta, un algoritmo de resolución de la incertidumbre contaría con dos soluciones alternativas. Por un lado considerar que la varianza equivale al centro entre los extremos (según estas últimas deducciones), por otro, que corresponde a casi la veintava parte de la amplitud al cuadrado. La decisión entre ambos resultados posibles será:

$$\sigma^2 = \max \left[\frac{\max - \min}{2}, \frac{(\max - \min)^2}{19.5} \right]$$

recurriendo a una estrategia relativamente conservadora, o bien:

$$\sigma^2 = \max \left[\frac{\max - \min}{2}, \frac{(\max - \min)^2}{12} \right]$$

en el caso de una actitud más claramente conservadora.

Máxima incertidumbre

Sin duda alguna es la situación más conflictiva, puesto que el investigador no conoce el valor aproximado de la media poblacional, ni su varianza, ni la distribución probable, ni los extremos esperables. En tal caso, nuestra estrategia consistirá en basar el valor arbitrario de la varianza poblacional en alguna medida cuya cuantía sea decidida subjetivamente por el investigador, de tal forma que la corrección de la situación, tras operar con los datos de la muestra, podrá consistir en variar la medida subjetiva en relación a la estimación objetiva de la varianza poblacional y su discrepancia con el valor considerado *a priori*.

El error de precisión (Manzano, 1997b) que el investigador considera, como variable previa a la decisión del tamaño de la muestra, será razonablemente dependiente del valor que se suponga para la medida objetivo. Así, si el valor obtenido con los datos de la muestra es de cuantía 5, carece de sentido un error de precisión con valor ± 10 . Igualmente, si la media tiene el valor 500, un error de precisión de cuantía 1 será juzgado unánimemente como minúsculo en exceso. Fuera de los extremos mencionados, es razonable pensar que la decisión sobre el valor que se considerará del error de precisión, esté en alguna medida dependiente de la cuantía

supuesta para la función poblacional cuya estimación justifica el estudio. Todo más, sabiendo que se trata de una decisión arbitraria y subjetiva (Silva, 1993).

En el caso de la estimación de medias, tal circunstancia puede simbolizarse mediante la expresión:

$$e_p = \frac{\mu}{h}$$

Este proceder, si bien no cuenta con un antecedente directo, sí viene apoyado por las reflexiones de Silva (op. cit.), según el cual, considerar $p=0.5$ cuando se desconoce la varianza poblacional en la estimación de proporciones con el objetivo de obtener un n máximo (conservador), es un error muy extendido que ignora la relación existente entre el error de precisión y el valor considerado a priori para el parámetro, puesto que ambas variables se mueven en el mismo sentido.

Una decisión esperable es tomar el valor $h=10$, de tal forma que el error de precisión sea la décima parte del valor considerado para la media poblacional. Así, si se considera que la media de edad de la población es aproximadamente 50, un error de precisión razonable es ± 5 . Aumentar la precisión implicará, por tanto, incrementar el valor de h hasta una cuantía satisfactoria para el investigador.

Aún así, en la situación que justifica el presente subapartado, el investigador no puede estimar un valor esperable para μ , por lo que la expresión anterior no es operativa. Sin embargo, nos resultará útil para las reflexiones que siguen.

Tras la recogida de datos, se comprueba la discrepancia entre la varianza considerada a priori (como medida procedimental), que simbolizaremos con σ_1^2 , y la estimada objetivamente a partir de las funciones muestrales pertinentes, que simbolizaremos con σ_2^2 . Caben, pues, tres posibles resultados, en términos del sentido de la discrepancia:

1. $\sigma_1^2 > \sigma_2^2$ La varianza poblacional fue sobreestimada, llevando a un tamaño muestral mayor de lo que hubiera resultado *óptimo*. Lo más esperable será reducir el error de precisión (o el tamaño de efecto), manteniendo el resto de las variables de cuantía decidida por el investigador (riesgos α y β).

2. $\sigma_1^2 < \sigma_2^2$ La varianza poblacional fue subestimada, llevando a un tamaño muestral menor de lo que hubiera resultado *óptimo*. Lo más esperable será aumentar el error de precisión (o el tamaño de efecto), manteniendo el resto de las variables de cuantía decidida por el investigador (riesgos α y β).

3. Ambos valores coinciden.

De las situaciones anteriores se observa con claridad la relación existente entre la varianza poblacional y el error de precisión, en el sentido de que las variaciones en una llevan a modificaciones en el otro. En definitiva, si consideramos tal relación expresada en términos

de una constante k:

$$\sigma^2 = k \times e_p \quad \text{Y} \quad k = \frac{\sigma^2}{e_p}$$

Por otro lado, el error de precisión había sido definido en términos del valor estimado para la medida objetivo. En el caso de la estimación de una media aritmética:

$$k = \frac{\sigma^2}{e_p} = \frac{h \sigma^2}{\mu}$$

Para resolver la expresión anterior de tal forma que sea útil a los contextos de máxima incertidumbre, utilizaremos de nuevo la analogía con el contexto de estimación de proporciones. En tal caso, interesa un valor máximo para k, de tal forma que se llegue a una estimación a priori de σ^2 con actitud conservadora.

$$k = \frac{h \pi (1 + \pi)}{\pi} \quad \text{Y} \quad \max(k) = h [1 + \min(\pi)] = h$$

Necesariamente, el investigador debe pronunciarse acerca de un valor para el error de precisión, aún sin conocimiento alguno acerca de cualquier característica poblacional. Si se conceptualiza éste como una porción de la medida objetivo, la estimación de la varianza como medida procedimental puede seguir las deducciones realizadas en este subapartado.

Un valor razonable para h es 10 (indirectamente, en Silva, 1993:147). Así, en el caso de $\pi = 0.5$, se obtendrían intervalos de estimación similares a (45%, 55%). En el caso de $\pi = 0.1$, el intervalo sería aproximadamente (9%, 11%). Valores mayores de h , llevarán a intervalos más reducidos, a valores estimados para σ^2 mayores y, consecuentemente, mayores tamaños de muestra. Así, la consideración de $h=20$, lleva a intervalos similares a (47.5%, 52.5%) en el caso de $\pi = 0.5$ y (9.5%, 10.5%) cuando $\pi = 0.1$.

En definitiva, es razonable estimar la varianza poblacional como 10 ó 20 veces el error de precisión considerado, en función de la actitud conservadora del investigador y, siempre y cuando se carezca de cualquier tipo de información objetiva acerca de alguna característica poblacional.

Otras medidas procedimentales

Correlación entre conglomerados δ

La correlación entre conglomerados es un índice que permite poner la ecuación del tamaño de muestra en función de la varianza de la población. De otro modo, el cálculo del número de unidades a seleccionar se encontraría en función de la varianza entre conglomerados y dentro conglomerados, aspectos éstos difíciles de acotar por el investigador. No obstante, la

estrategia si bien elimina algunas funciones de difícil valoración, añade una nueva incertidumbre: el valor para δ .

Para encontrar un valor satisfactorio de δ , pueden ponerse en marcha los mismos procedimientos mencionados en el caso de las varianzas: revisión de la literatura, puesta en marcha de un estudio piloto y partición de la muestra en premuestra y resto o en muestra inicial y suplemento. Además, existen algunas estrategias específicas para el caso de δ . Así, en el trabajo de Smith (1938) se propone la expresión

$$\sigma_e^2 = \sigma^2 G^\alpha$$

mientras que Jessen (1942) propone

$$\sigma_d^2 = A G^g$$

Ambos (citados en Mirás, 1986), se fundamentan en el hecho de que el tamaño medio de conglomerado tiene una clara influencia sobre las varianzas. Si G fuera todo lo más grande posible, coincidiría con la población, con lo que existiría un único conglomerado y toda la varianza sería *dentro*. En el otro extremo, el tamaño más pequeño posible para un conglomerado es de una única unidad, en cuyo caso toda la variación sería *entre*. No obstante, las soluciones de Smith y Jessen son relativas, por cuanto que el desconocimiento de las varianzas es sustituido por el desconocimiento de las variables que proponen (α , A y g), cuyos valores son despejados mediante la práctica o experiencia en el campo de aplicación específico en el que se realice el muestreo. Al respecto, hay que señalar que ambos autores trabajaron en el contexto de la investigación en agricultura.

Por último, queda la estrategia de suponer algún valor operativo y razonable para δ que permita continuar con el procedimiento de decisión sobre n , aún en el caso de que el investigador

carezca de indicaciones empíricas sobre tal índice. Para ello, es previo acotar su rango de variación.

1. En uno de los extremos, se encuentra la situación en la que σ_e^2 (variación entre conglomerados) = 0. En tal caso, toda la variación existente se encuentra dentro de los conglomerados, ya que éstos son réplicas unos de otros, al menos en cuanto al valor promedio de la medida objetivo. Es la situación en la que se requiere el mínimo tamaño de muestra: un único conglomerado, puesto que provee del mismo valor que la población en su conjunto.

$$\sigma^2 = \sigma_e^2 + \sigma_d^2$$

$$\delta = \frac{\sigma_e^2 + \frac{\sigma_d^2}{G+1}}{\sigma^2} = \frac{\sigma_e^2 + \frac{\sigma_d^2}{G+1}}{(\sigma_e^2 + \sigma_d^2)} = \frac{1}{G+1}$$

2. En el otro polo del intervalo se encuentra la situación en la que σ_d^2 (variación dentro conglomerados) = 0. Por tanto, toda la variación se encuentra entre los conglomerados. Éstos constituyen repeticiones del mismo valor, dentro de sí. Es la situación que exige mayor tamaño de muestra. De hecho, una única unidad es una buena representación del conglomerado al que pertenece, mientras que el procedimiento exigirá seleccionar tantos conglomerados en el modelo de selección de conglomerados, como unidades elementales en el modelo aleatorio simple.

$$\sigma^2 = \sigma_e^2 + \sigma_d^2$$

$$\delta = \frac{\sigma_e^2 + \frac{\sigma_d^2}{G+1}}{\sigma^2} = \frac{\sigma_e^2}{\sigma^2} = 1$$

Así pues y considerando que el conglomerado de menor tamaño en el que todavía pueda calcularse variación es $G=2$, el intervalo de variación para δ será:

$$\delta \in \left[\frac{1}{G-1}, \frac{1}{G} \right] = \left[\frac{1}{2-1}, \frac{1}{2} \right] = \left[1, \frac{1}{2} \right]$$

Una vía de solución en las situaciones donde el investigador desconoce totalmente el valor de δ es tomar el centro del intervalo, es decir:

$$\tilde{\delta} = \frac{1 + \frac{1}{G+1}}{2} = \frac{G+2}{2G+2}$$

Lo usual es que el valor δ sea positivo (Mirás, 1986), lo que concuerda con el valor sugerido para $\tilde{\delta}$, que nunca es negativo y que únicamente se anula para el caso inesperado de $G=2$.

Una aproximación alternativa para la concreción de un valor δ consiste en expresar una de las dos varianzas implicadas en función de la otra mediante una constante k :

$$\sigma_d^2 = k \sigma_e^2$$

En tal caso:

$$\sigma^2 = \sigma_d^2 + \sigma_e^2 (k+1) \quad \text{Y} \quad \sigma_e^2 = \frac{\sigma^2}{k+1}$$

Con ello, la expresión δ quedaría únicamente en términos del tamaño medio de conglomerado y de la constante k :

$$\delta = \frac{\sigma_e^2 + \frac{\sigma_d^2}{G+1}}{\sigma^2} = \frac{\frac{\sigma^2}{k+1} + \frac{k\sigma^2}{(k+1)(G+1)}}{\sigma^2} = \frac{G + (k+1)}{(G+1)(k+1)}$$

En el caso particular:

$$k = 1 \quad \text{Y} \quad \sigma_e^2 = \sigma_d^2$$

El valor estimado para δ será:

$$\delta = \frac{G+2}{(G+1)^2} = \frac{G+2}{2G+2}$$

Por lo que el desconocimiento de la relación entre las varianzas entre y dentro conglomerados, de sus valores y/o de la cuantía para δ , lleva a una misma estimación a priori basada tanto en la suposición de igualdad de varianzas como en el centro del recorrido de valores posibles para δ .

Correlación entre grupos relacionados ρ

Cuando el contexto de inferencia implica la comparación de medias en grupos relacionados, el tamaño de muestra requiere el conocimiento de una medida de covariación entre los grupos que se comparan, como es el caso de la correlación lineal de Pearson.

El problema, pues, es estimar tal índice. La solución pasa por cualquiera de las consideradas en subapartados anteriores: revisión de la literatura, estudio piloto, premuestra o partición en muestra inicial y suplemento. Además existe el recurso de considerar los valores posibles del índice y escoger uno de ellos como el más idóneo en las situaciones de máxima incertidumbre.

El índice se encuentra acotado por

$$-1 \leq \rho \leq 1$$

Varios argumentos recomiendan la solución $\rho = 0$ en las situaciones donde no exista ninguna información que sugiera otra cosa. Principalmente, cero es el punto intermedio entre las cotas del índice (lo que implica, si la distribución de posibles valores es simétrica, un error medio y mediano mínimo). Además cero es el valor que no indica ninguna dirección concreta de relación. Si no existe ninguna información que apunte hacia una relación positiva o negativa, una solución de compromiso es mantener que no existe relación.

Corrección del tamaño de muestra en función del tamaño piloto

Cuando se realiza un estudio piloto y se requiere una medida de la varianza poblacional, lo usual es realizar una estimación puntual de ésta a partir del valor de variación obtenida en el conjunto de datos del estudio piloto. En general, las medidas procedimentales se estiman puntualmente, mientras que las medidas objetivo se someten a una estimación por intervalo. Tal procedimiento es razonable puesto que la pretensión de una estimación por intervalo para cada

medida requerida lleva a una progresión infinita de estimaciones.

Pongamos por caso que se requiere un valor para la varianza poblacional en el cálculo del tamaño de la muestra, en un muestreo aleatorio simple para la estimación de la media aritmética en una población de gran tamaño. La estimación de la media se realizará mediante el intervalo

$$\mu \in (\bar{x} \pm Z_{\alpha} \sigma_{\bar{x}})_{1-\alpha} \quad \text{Y} \quad \mu \in (\bar{x} \pm Z_{\alpha} \sigma_{\bar{x}}, \bar{x} \pm Z_{\alpha} \sigma_{\bar{x}})_{1-\alpha}$$

Se necesita calcular el tamaño adecuado para la muestra y, para cubrir tal necesidad además de otras cuestiones, se recurre a realizar un estudio piloto. El número de unidades seleccionadas para esta preinvestigación es n_p y se obtiene una desviación tipo de cuantía S_p . Considerando el modelo de muestreo utilizado, un estimador insesgado de la varianza poblacional es la cuasivarianza de la muestra. Si se realiza una estimación puntual, se considerará:

$$\sigma^2 = S^2 \frac{\sqrt{n_p}}{\sqrt{n_p - 1}} \quad \hat{S}$$

Es la expresión anterior la que se utiliza como elemento para calcular el valor del error tipo en la distribución muestral de medias. Pero supongamos que se realiza una estimación por intervalo de la varianza poblacional, en tal caso:

$$\min(\sigma) = \hat{S} \pm Z_{\alpha} \sigma_{\hat{S}} \quad \text{Y} \quad \max(\sigma) = \hat{S} \pm Z_{\alpha} \sigma_{\hat{S}}$$

Con ello, la estimación de la media deberá modificarse con los extremos ampliados según las fluctuaciones que corresponden a la estimación de la varianza. Considerando ambos extremos

esperables para σ , el error tipo fluctuará, a su vez, entre los valores extremos:

$$\min(\sigma_{\bar{x}}) = \frac{\hat{S} \cdot Z_{\alpha} \sigma_{\hat{S}}}{\sqrt{n}} \quad \text{y} \quad \max(\sigma_{\bar{x}}) = \frac{\hat{S} \cdot Z_{\alpha} \sigma_{\hat{S}}}{\sqrt{n}}$$

Por lo que la forma de la estimación por intervalo de la media será:

$$\mu \in \left(\bar{x} \pm Z_{\alpha} \frac{\hat{S} \cdot Z_{\alpha} \sigma_{\hat{S}}}{\sqrt{n}}, \bar{x} \pm Z_{\alpha} \frac{\hat{S} \cdot Z_{\alpha} \sigma_{\hat{S}}}{\sqrt{n}} \right)_{1-\alpha}$$

El intervalo de estimación ha quedado sensiblemente complicado, pero el problema de evitar las estimaciones puntuales no se ha eliminado. La expresión contiene ahora una nueva medida: el error tipo de la distribución muestral de cuasidesviaciones típicas que, a su vez, estará en función de medidas procedimentales que habrán de estimarse, volviendo a generarse un dilema entre estimación puntual o por intervalo.

Luego, concluimos tal y como se inició el presente apartado: las estimaciones de las medidas procedimentales deben realizarse por punto, dejando las estimaciones por intervalo para las medidas objetivos.

No obstante, la salida es incómoda por cuanto resulta totalmente insensible al poder de estimación en el estudio piloto. Así, se concede la misma relevancia a una estimación de varianzas realizada a partir de un estudio piloto con $n_p=10$ que con $n_p=100$. Si ambos estudios se han realizado con las mismas garantías de control, los resultados provenientes de la segunda preinvestigación son, cuando menos, más creíbles. Tales reflexiones sobre el tamaño de muestra en el estudio piloto son extensibles a las situaciones en las que el investigador toma valores para

las medidas procedimentales que requiere, de otras investigaciones anteriores (con tamaños de muestra explícitos).

Ante esta situación, Shiffler y Adams (1987) proponen un método para corregir el tamaño final de la muestra n en función del tamaño del estudio piloto n_p . De esta forma, conforme n_p disminuye, n aumenta puesto que, razonablemente, la estimación de una varianza poblacional a partir de una muestra más pequeña es menos fiable y está sujeta a más fluctuaciones. Mediante un desarrollo que contempla distribuciones muestrales normales, los autores proponen una tabla de corrección de n , de la que la figura 5 muestra una información parcial.

n_p	C
3	1.443
4	1.267
5	1.192
6	1.149
8	1.089
10	1.071
12	1.064
15	1.049
20	1.036
40	1.017
60	1.011

Figura 5: *Tabla de relación entre el tamaño de muestra en el estudio piloto y el factor multiplicador para el tamaño final de muestra.*

El factor de corrección C es la constante por la que debe multiplicarse al tamaño inicial calculado n' , para conseguir el tamaño final n . Por otro lado, para tamaños $n_p > 60$, el factor de corrección es prácticamente la unidad, con lo que el tamaño final de la muestra no debe sufrir variación.

Bibliografía

Azorín, F.; Sánchez Crespo, J.L. (1986). *Métodos y aplicaciones del muestreo*. Colección Alianza Universidad Textos. Madrid: Alianza Editorial.

Barnett, V. (1974). *Elements of sampling theory*. London: The English University Press.

Cochran, W. G. (1976). *Técnicas de muestreo*. México: Compañía Editorial Continental S.A..

Czaja, R. y Blair, J. (1996). *Designing surveys. A guide to decisions and procedures*. Thousand Oaks, California: Pine Forge Press.

Feigl, P. (1978). A graphical aid for determining sample size when comparing two independent proportions. *Biometrics*, 34, 111-122.

Fink, A. y Kosecoff, J. (1989). *How to conduct surveys: a step-by-step guide*. Beverly Hills: Sage.

Gillett, R. (1989). Confidence interval construction by Stein's method: a practical and economical approach to sample size determination. *Journal of marketing research*, 26, Mayo, 237-240.

Gutiérrez Cabría, S. (1978). *Bioestadística*. Madrid: Tebar Flores.

Hedges, B. (1980). Sampling. En G. Hoinville y R. Jowell (Eds.). *Survey research practice* (pp. 55-89). London: Heinemann Educational Books.

Henry, G.T. (1990). *Practical sampling*. Applied Social Research Methods Series, vol 21. Newbury Park, California: Sage Publications.

Kalton, G. (1987). *Introduction to survey sampling*. SAGE University Paper. Beverly Hills: SAGE Publications Inc.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

McCall, Ch. H. Jr. (1982). *Sampling and statistics Handbook for research*. Ames, Iowa:

The Iowa State University Press.

Mirás, J. (1986). *Elementos de muestreo para poblaciones finitas*. Madrid: Instituto Nacional de Estadística.

Salgado, J.A. (1990). La práctica del muestreo. En E. Ortega (Eds.). *Manual de investigación comercial* (pp. 344-377). Madrid: Pirámide.

Shiffer, R.E. y Adams, A.J. (1987). A correction for biasing effects of pilot sample size on sample size determination. *Journal of marketing research*, 24, August, 319-321.

Silva, L.C. (1993). *Muestreo para la investigación en ciencias de la salud*. Madrid: Diaz de Santos.

Sudman, S. (1983). Applied sampling. En P.H. Rossi, J.P. Wright y A.B. Anderson (Eds.). *Handbook of survey research* (pp. 145-197). Quantitative Studies in Social Relations. Orlando, Florida: Academic Press.

Teijeiro, F. (1990). Técnicas de muestreo. En E. Ortega (Eds.). *Manual de investigación comercial* (pp. 312-343). Madrid: Pirámide.

Wilburn, A.J. (1984). *Practical statistical sampling for auditors*. Nueva York: Marcel Dekker.