ESTADÍSTICA ESPAÑOLA
Vol. 44, Núm. 151, 2002, págs. 307 a 335

e- Encuestas Probabilísticas II. Los Métodos de Muestreo Probabilístico

por
ANA M. MUÑOZ REYES
M. DOLORES JIMÉNEZ GAMERO
JOAQUÍN MUÑOZ GARCÍA
RAFAEL PINO MEJÍAS
Departamento de Estadística e Investigación Operativa. Facultad de Matemáticas, Universidad de Sevilla

RESUMEN

En este trabajo se aborda fundamentalmente el estudio de las encuestas que utilizan la herramienta de Internet para su realización. En concreto su objetivo se centra en el planteamiento y desarrollo de diseños muestrales probabilísticos que permitan realizar encuestas desde la World Wide Web con el rigor necesario para poder inferir los resultados obtenidos a la población objeto de estudio, con determinada fiabilidad.

Palabras claves: encuesta, Internet, e-encuesta, diseño muestral probabilístico.

Clasificación AMS: 62D05, 62P99

1. INTRODUCCIÓN

La actual y progresiva incorporación de Internet al campo de las encuestas ha dado lugar a la introducción de nuevos conceptos así como a la actualización y adaptación de técnicas ya existentes. Dependiendo de la componente de la Red (correo electrónico, transferencia de ficheros (FTP), World Wide Web,...) que se utilice en la realización de la encuesta, habrán de considerarse unos u otros aspectos de las encuestas por muestreo.

Así, en el caso del correo electrónico la situación y problemática que se presenta es similar a las clásicas encuestas postales. Sin embargo, no ocurre lo mismo con el uso de otras componentes como puede ser la utilización de la World Wide Web (www) cuyas características particulares hacen que, para seleccionar una muestra probabilística sobre la que realizar la encuesta, sea necesario introducir nuevos conceptos y modificar o adaptar los métodos estadísticos disponibles.

Las definiciones de e-encuesta, e-población, así como la clasificación de ésta última en población saturada/no-saturada en Internet, con sus correspondientes subclasificaciones, como queda recogido en Cubiles, Muñoz, Pascual y Muñoz (2002) se utilizarán en este trabajo como punto de partida, desarrollando en cada caso los métodos de muestreo probabilístico adecuados a cada situación.

En primer lugar, en la sección 2, se estudia la selección de unidades de muestreo, distinguiendo según se realice dentro de la Red o de forma externa a la misma.

En las siguientes secciones de este trabajo se desarrollan los métodos de muestreo probabilístico para los distintos tipos de población objetivo definidos en Cubiles et al. (2002). Así, la Población Saturada en Internet queda desglosada en cuatro tipos. En la sección 3 se aborda la Población Audiencia, distinguiendo según los elementos de muestreo a considerar sean las visitas o los visitantes. En el primer caso, la población sobre la que se realizarán los procesos de inferencia queda fijada por los elementos que acceden a un site determinado de antemano durante el periodo de realización de la encuesta, identificándose cada uno de ellos como una visita. Por otro lado, cada internauta que accede alguna vez en dicho periodo de tiempo al site se considera un visitante. El estudio de los métodos de muestreo probabilístico a aplicar en cada caso caso se desarrollan en las secciones 3.1 y 3.2, respectivamente.

La sección 4 presenta las particularidades correspondientes al estudio de una Población Precisada.

La Población Internet y la Población Internet Especial se recogen en las secciones 5 y 6, respectivamente. En ambos casos se desarrolla un muestreo probabilís-

tico bietápico definiendo como unidades primarias las sites que recubren la e-población y como unidades secundarias las visitas correspondientes a cada site.

Con una Población No Saturada en Internet será necesario utilizar un marco y aplicar un diseño muestral adecuado a la población definida, independientemente de los medios que dispongan sus elementos para navegar por Internet. La accesibilidad de un elemento a la Red será a lo sumo utilizada para aplicar un diseño con un marco dual. Esto queda recogido en la Sección 7.

Este trabajo concluye con una sección dedicada a tratar la disminución de los errores ajenos al muestreo en distintos aspectos y etapas de la realización de la encuesta.

2. MUESTREO E INTERNET

Al hacer encuestas con entrevistas en la Red a través de la www, pueden considerarse dos situaciones: una en que las unidades elementales de muestreo son seleccionadas de forma externa a la Red, en cuyo caso pueden incluirse poblaciones objetivos como las precisadas o las no saturadas en Internet según lo definido en Cubiles et al.(2002), y la otra situación se presenta cuando la selección de las unidades que formarán parte de la muestra se realiza desde la propia Red, lo que se hará mediante la interceptación de los internautas cuando éstos estén navegando en Internet. En esta última pueden incluirse poblaciones objetivos como la referida a la audiencia a un “site" o red de “sites”, la población Internet y la población Internet especial según Cubiles et al.(2002).

2.1 Selección Externa a la Red

En este caso, las unidades de muestreo y su número serán precisados por el marco correspondiente a la población objetivo de la encuesta, y en dicho caso las informaciones que proporcione el marco serán las clásicas a cualquier encuesta a la que se ha de añadir cierta valoración sobre las posibilidades de realizar la entrevista a través de la Red.

En este caso las unidades de muestreo se seleccionarán según procedimientos adecuados a las condiciones que se tengan sobre la encuesta a realizar y donde las unidades son contactadas a voluntad del entrevistador y podrán aplicarse los diseños muestrales y los esquemas de muestreo tradicionales.

2.2 Selección en la Red

En este caso las unidades elementales de muestreo que forman parte del marco y que identifican a la población objetivo serán los “surfistas” que existan en la Red durante el período de realización de la encuesta, y para los que se considera que los accesos a los “sites” de un “surfista” a otro son independientes, incluso los del mismo “surfista” a diferentes “sites” o en su reincidencia a un mismo “site”. Por tanto la población objetivo estará condicionada por las características de los elementos que hayan accedido a la Red durante el tiempo de realización de la encuesta y el número de unidades elementales de la población de referencia será conocido de una forma precisa al finalizar el proceso de muestreo, aunque como se verá más adelante pueden darse situaciones en que el número de elementos de la población puede ser fijado con antelación a la realización de la encuesta.

Los diseños muestrales a aplicar sobre la población de “surfistas", serán generalmente el muestreo aleatorio simple sin reemplazamiento o el muestreo sistemático. Estos diseños muestrales pueden precisarse utilizando diseños estratificados, donde las variables estratificadoras pueden ser algunas como las que se indican a continuación:

Horas del Día: Los “surfistas” pueden presentar unas características y una variabilidad muy diferente según el horario de acceso a Internet. Pueden tomarse estratos como, horario laboral, el horario de tarifa plana o el horario nocturno frente a la variabilidad de los husos horarios para la población global, etc.

Días de la Semana: Los “surfistas” pueden tener también características y variabilidades dispares, según los días de la semana en que éstas accedan al “site”, así pueden considerarse los fines de semana y días festivos frente al resto de la semana, etc.

Páginas de la Web: Las páginas que se visiten en la “Red” también puede dar lugar a caracterizar o distinguir a los elementos de la e-población.

La estratificación de la población dependerá de los objetivos que se planteen en la encuesta y de la información de que se disponga en el marco. Al definir los estratos, éstos deben cumplir la condición de tener intersección vacía. Por ejemplo, en el caso de los días de la semana, pueden definirse tres estratos como el debido a los “surfistas” que acceden a Internet los fines de semana y días festivos únicamente, los que lo hacen durante el resto de la semana y los que acceden a la Red cualquier día de la semana.

2.2.1 Métodos de selección

Los métodos de selección a emplear en la Red tendrán que ser métodos secuenciales que se aplican al acceder los internautas a determinados "sites" de la Red, es decir, las unidades de muestreo son seleccionadas una a una según los criterios de elegibilidad que se impongan sobre las unidades elementales.

Fan, Müller and Rezucha (1962), dan un conjunto de métodos de selección secuencial de forma que las muestras que se obtengan mantengan, en el caso de aplicar muestreos probabilísticos, las mismas probabilidades que las muestras obtenidas por métodos no secuenciales, tratándolos tanto en el caso de conocer el número total de unidades elementales sobre la que se está muestreando, como en el caso de no conocer dicho número. Muchos de estos métodos se han ido perfeccionando y han ido apareciendo otros alternativos. En esta línea pueden citarse, entre otros, el que se recoge en McLeod and Bellhouse (1983), quienes dan un procedimiento secuencial aplicable tanto para cuando el número total de unidades de muestreo es desconocido como cuando no lo es, o el de Bissell (1986), que propone un método de selección aleatoria ordenada sin reemplazamiento para cuando se conoce el número total de unidades de muestreo, este método es mejorado por Pinkham (1987), quien le añade el caso de que el tamaño poblacional sea desconocido.

Todos los métodos de selección suelen ir acompañados del correspondiente software para su aplicación. Por tanto, resulta necesario discutir en muchos casos los algoritmos que lo implementan, como hacen Bellhouse and Koulperger (1991), que analizan la eficiencia en la ejecución desde el punto de vista de la velocidad y el espacio de almacenamiento que necesitan.

A continuación se desarrollaran los métodos de muestreo probabilístico para los tipos de población objetivo definidas en Cubiles et al.(2002) y con los fundamentos de muestreo descritos previamente.

3. MUESTREO PROBABILÍSTICO PARA LA POBLACIÓN AUDIENCIA

Al estudiar el marco de la población audiencia a un "site" pueden generarse dos poblaciones: la población audiencia-visitas y la población audiencia-visitantes. En este caso, y con el fin de que la condición de elegibilidad de las unidades de muestreo sea determinada de una forma precisa, se considerarán como unidades de muestreo la de los "surfistas" de Internet que accedan al site, denominándolos "visitas al "site''. Por tanto se contactará con los internautas para que respondan la encuesta mediante su interceptación cuando accedan al "site".

3.1 Muestreo Probabilístico, Población Audiencia - Visitas

En este muestreo se supondrá que el acceso de las visitas al “site” son independientes unas de otras, incluso las que pueda realizar un mismo internauta. La consideración de visitas desde el punto de vista de la audiencia puede ser sometida a ciertas restricciones, como pueden ser, fijar algún tiempo de permanencia en el “site”, visitar determinadas páginas o determinado número de páginas del “web site”, etc., y lo aconsejable es que los criterios que se impongan para considerar un acceso al “site” como visita, sean objetivos para observarlos de forma externa al internauta, con el fin de que la interceptación de los elementos que formen parte de la muestra se realice únicamente a los “surfistas” que cumplan de forma exacta la condición de elegibilidad exigida para ser considerados visitas.

3.1.1 Tamaño de la Población

La población audiencia visitas, que representaremos mediante $P_v = \{v_1, v_2, ..., v_{N_v}\}$ como población sobre la que se realizarán los procesos de inferencia correspondiente, queda fijada por los elementos que acceden al “site” durante el período de realización de la encuesta. En este caso, el tamaño de la población que se estudia y el número de unidades de muestreo que se genera desde el marco es el mismo, es más, en este caso se tiene un marco perfecto salvo las imperfecciones que surjan de la realización de la encuesta.

Aunque el tamaño de la población, N_v, sólo será conocido de manera exacta cuando finalice el periodo de realización de la encuesta, puede disponerse de información previa sobre el número de visitas al “site" en periodos de tiempo próximos a la realización de la encuesta, lo cual es posible por el software que se utiliza para gestionar la “web". Esto nos permitirá conocer previamente de manera aproximada N_v.

Otra forma de proceder es obtener el tamaño de la población fijando la fracción de muestreo para un número fijo de visitas a seleccionar de la población.

Por tanto, se tiene de una forma genérica, que todos los estimadores o conclusiones que se obtengan estarán referidas al número de elementos de la población de visita que se haya fijado u obtenido, y por supuesto al tiempo de realización de la encuesta.

Planteamientos similares hay que hacerse, en el caso de que se aplique un diseño muestral estratificado, sobre el tamaño o número de unidades en los distintos estratos.

3.1.2 Diseños muestrales

Los diseños muestrales que se aplicarán a la población audiencia - visitas, serán alguno de los dos indicados en la sección anterior, es decir, el muestreo aleatorio simple sin reemplazamiento o el muestreo sistemático, y por supuesto puede plantearse un muestreo estratificado con variables estratificadoras similares a las indicadas u otras para las que sea posible construir los estratos.

Los procedimientos de selección a utilizar para la interceptación variarán según se suponga conocido o no, el número de elementos de la población. En el primer caso es posible aplicar un conjunto de métodos de selección más amplio que en el segundo.

Si se aplica un muestreo estratificado, el problema de la afijación muestral en cada estrato dependerá de la información que se tenga, pero en cualquier caso la situación más simple y que siempre será aplicable, es la de la afijación proporcional, delimitada según se suponga conocido el número de unidades existentes en cada estrato o no. En esta última situación puede utilizarse para realizar la afijación una variable auxiliar como puede ser las visitas realizadas a cada uno de los estratos lo que siempre es posible disponer.

Las probabilidades de que las visitas pertenezcan a la muestra en los métodos de muestreo indicado, los estimadores que se utilizan para estimar determinadas funciones paramétricas y sus respectivas varianza y estimaciones de éstas, son recogidos en textos clásicos como los de Azorín y Sánchez - Crespo (1986) y Särndal, Swensson and Wretman (1992).

3.1.3 Tamaño de la muestra

En todos los diseños muestrales que se apliquen sobre la población de visitas, se considerará inicialmente fijado el tamaño de la muestra n_v. Dicho valor depende generalmente de la precisión que se desee tener en los procesos de inferencia y del número de elementos de la población. Si este número resulta inicialmente desconocido, podría adoptarse la solución de fijar una fracción de muestreo pequeña. Esto no supone ningún tipo de restricción práctica, ya que generalmente los "sites" en los que se esté interesado en estudiar su audiencia tendrán generalmente la peculiaridad de tener un número elevado de visitas. En caso contrario puede ocurrir que se observen todos los elementos de la población.

3.2 Muestreo Probabilístico, Población Audiencia – Visitantes

Un visitante a un "site" o conjunto de "sites" fijados es un internauta que accede alguna vez durante el período de tiempo que se realice la encuesta a dicho "site" o conjunto de "sites". También puede considerarse con referencia a las visitas, indi-

cando que un visitante es una visita, sin considerar sus posibles recurrencias o accesos repetidos al "site" o conjunto de "sites". En lo que sigue se considerará como referencia un único "web site".

Sobre la población de visitantes, que denotaremos mediante $P_V = \{V_1, V_2, ..., V_{N_v}\}$ $(N_V \leq N_v)$ pueden derivarse procesos de inferencia estadística desde la desagregación de e-encuestas probabilísticas realizadas sobre la población de visitas. En estos casos, y para realizar estimaciones, se utilizarán estimadores de razón, ya que el número de visitantes en dichas e-encuestas es inicialmente desconocido y por tanto hay que estimarlo. Pero esta forma de proceder suele conducir a que las estimaciones que se obtengan no satisfagan determinados niveles de precisión deseables en toda encuesta. Por tanto, el método de muestreo aconsejable debería fijar un tamaño n_V para la muestra de visitantes a obtener de la población de visitas, y que se representará por m_V. El tamaño n_{V_V} se fijará de forma que satisfaga las restricciones impuestas a los errores de muestreo.

El marco a utilizar para estudiar la población de visitantes se obtendrá a partir del marco de la población de visitas, ya que se aplicará el método de entrevista por interceptación, definiendo una relación de equivalencia en la que dos visitas están relacionadas si corresponden al mismo visitante, lo que generará las clases de equivalencia

$$C_j = \{v_k \in P_v \mid v_k \rightarrow V_j \quad k = 1,2,...,N_v\}, \quad j = 1,2,...,N_V.$$

Por tanto, $\#C_j$ representa el número de visitas que realiza el visitante o internauta j-ésimo, para j=1,2,...,N_v, que se interpretará como el grado de multiplicidad que presenta el elemento j-ésimo de la población de visitantes, al considerar para su estudio las unidades de muestreo procedente del marco que genera la población de visitas. Por tanto, para estudiar la población de visitantes se utilizarán los mismos diseños muestrales que en el caso de la población de visitas, con las correspondientes correcciones por multiplicidad del marco.

Existen muy diversos métodos para realizar correcciones en los diseños muestrales cuando se tienen marcos con multiplicidad, como puede verse en Lessler and Kalsbeek (1992). Uno de estos métodos, pondera las observaciones según el grado de multiplicidad que éstas tengan. Este método a su vez permite distintas situaciones,

1. Se conoce el valor de $\#C_j$, para los elementos de la muestra,

   a) por información previa

   b) aplicando métodos de selección secuencial de la muestra, como puede ser el de McLeod and Bellhouse (1983)

2. Se obtiene el valor de $\#C_j$, durante la entrevista (incluyendo en el cuestionario preguntas que permitan determinarlo).

Cuando no sea posible conocer $\#C_j$, ni siquiera para los elementos de la muestra, se utilizarán los cardinales de las clases de equivalencia que resultan de la muestra

$$C_j^{'} = \{v_k \in m_v |\ v_k \to V_j\} \quad j = 1,2,...,N_V,$$

donde el grado de multiplicidad en la muestra viene dado por $\#C_j^{'}$, $\forall$ j=1,2,...,N_v. Este valor puede obtenerse sin necesidad de incluir ninguna pregunta el cuestionario, bastará con tener la identificación que se propuso para los entrevistados.

3.2.1 Tiempo de Realización de la Encuesta.

En el muestreo probabilístico que se ha descrito para la audiencia de visitantes, se ha fijado la necesidad de obtener un número de visitantes de acuerdo a los niveles de precisión que se hayan fijado para el proceso de estimación. Esto conduce a que el tamaño de la muestra de visitas m_V, que se representará por n_V, resulte ser una variable aleatoria, para la que puede plantearse la cuestión de los posibles valores extremos a tomar, ya que ello afectará al tiempo de realización de la encuesta.

Sea m_V la muestra de visitas en la que se identifica una submuestra de visitantes m_V, a la que se le exige que tenga un tamaño n_V=n_{V_0}. Supuesto que los accesos al "site" de un "surfista" a otro, e incluso las reincidencias de un surfista al mismo "site" son independientes e idénticamente distribuidos, se tiene que

$$P\left[n_v = n_{V_0} + r\left(n_{V_0}, p_V\right)\right] = \left[\binom{n_{V_0}+r-2}{r} p_V^{n_{V_0}-2}(1-p_V)^r\right] p_V \qquad r = 0,1,2,.....$$

con p_V= Probabilidad de que un visitante del site realice al menos una visita durante el periodo de realización de la encuesta.

El estimador de máxima verosimilitud para la probabilidad de ser visitante viene dado por

$$\hat{p}_V = \frac{n_{V_0} - 1}{n_v - 1}$$

el cual no es insesgado, siendo un estimador insesgado para dicha probabilidad

$$\left(\hat{p}_V\right)' = \frac{n_{V_0} - 2}{n_v - 2}$$

tal como demuestra Kendall and Stuart (1969).

La función de distribución para la variable aleatoria número de visitas vendrá expresada de la siguiente forma

$$F_v(n-1) = P\left[n_v \leq n-1 \middle| n_{V_0} - 1, p_V\right] = p_V^{n_{V_0-1}} \sum_{r=0}^{n-n_{V_0}} \binom{n_{V_0} + r - 2}{r} (1-p_V)^r,$$

donde se ha tomado $(n\text{-}1)$ como referencia ya que la primera observación no se ha incluido. Morris (1963) demuestra que dicha función de distribución es idéntica a la probabilidad P_B de obtener al menos n_{V_0}-1 éxitos en una distribución binomial de parámetros $(n_v - 1, p_V)$

$$P_B = \sum_{x=n_{V_0}-1}^{n_v-1} \binom{n_v - 1}{x} p_V^x (1-p_V)^{n_v-1-x} = P\left[X \geq n_{V_0} - 1 \middle| n_v - 1, p_V\right]$$

Al ser el tamaño $n_V \geq n_{V0}$ y ser éste un valor elevado ya que debe adecuarse a los errores fijados, la probabilidad P_B puede aproximarse por la de una ley normal de parámetros $N((n_v - 1)p_{V,}(n_v - 1)p_V(1-p_V))$ lo que permitirá calcular una región de confianza para el tamaño de la muestra de visitas dada la probabilidad de ser visitante, p_V

$$1-\alpha \leq P\left[X \geq (n_{V_0} - 1) \middle| (n_v - 1), p_V\right] \Rightarrow$$

$$\Rightarrow (n_{V_0} - 1) - (n_v - 1)p_V \leq -z_{1-\alpha}\sqrt{(n_v - 1)p_V(1-p_V)}$$

siendo z_{1}-α el percentil de orden 1-α de la distribución $N(0,1)$ y donde α tomará, como es habitual, valores del orden de 0.10 o inferior. Cotas superiores para el tamaño de la muestra n_v según determinados niveles de confianza y para valores fijados de n_{V_0} y p_V, se muestran en la Tabla 1.

De ello se desprende que el proceso de encuestación siempre será finito y realizable para unos valores del tamaño de la muestra para la población visitantes fijado previamente.

4. MUESTREO PROBABILÍSTICO EN UNA POBLACIÓN PRECISADA

Para este tipo de poblaciones el proceso de entrevista no es del tipo de interceptación, ya que al disponer de un marco donde los elementos son precisados con su identificación, se procederá a diseñar el método de muestreo probabilístico adecuado para la información que se dispone e identificar los elementos que formarán parte de la muestra, tras lo cual se contactará con las unidades de muestreo para que respondan al cuestionario a través de la Red. El contacto con las unidades de muestreo puede realizarse por cualquier método como correos, teléfono, correo electrónico, etc., lo que dependerá de los objetivos, la temporalidad, la información auxiliar del marco, etc.

Tal y como afirma Cubiles et al. (2002) al estudiar el marco para este tipo de población, resulta necesario estimar la tasa de recubrimiento que tendrá la e-población respecto a la población precisada. La estimación de la tasa puede realizarse a partir de la información que se obtenga al contactar con las unidades de muestreo que forman la muestra.

5. MUESTREO PROBABILÍSTICO EN LA POBLACIÓN INTERNET

La Población Internet o población de internautas se representará por $P_I = \{I_1, .I_2..I_N\}$, y sus elementos serán precisados al fijar su comportamiento frente a la Red según los objetivos que se planteen para la encuesta.

Siguiendo la definición dada de encuesta, por Seco, A.; Olimpia, A. y Ramos, G. (2000), puede decirse que sobre la Población Internet se han hecho y se están haciendo muchas encuestas que utilizan la Red para hacer las correspondientes entrevistas. En el caso español, puede citarse como una de las más relevantes la que hace con cierta periodicidad la AIMC (Asociación para la Investigación de los Medios de Comunicación) a la población de “usuarios de Internet que visitan“ sites españoles’’ y que puede observarse en http: //www.aimc.es/.

Como se indica en la ficha técnica de la citada encuesta, ésta no está fundamentada en un muestreo probabilístico, e indica mediante ciertas matizaciones las reservas a tener cuando se realicen conclusiones de los resultados que obtienen. A continuación se recogen algunas de esas matizaciones

“La muestra final no es el resultado de una selección realizada desde la administración del estudio, sino que simplemente se incluye a aquellas personas que voluntariamente han aceptado y decidido colaborar (muestra autoseleccionada)”

"La dirección del sesgo que se obtiene con el procedimiento utilizado es conocido y ha sido suficientemente estudiado. La muestra sobrerepresenta a los internautas que hacen un uso más intenso de la red, los más experimentados, etc."

De igual modo pueden citarse otras encuestas realizadas en otros países, en particular en Estados Unidos, donde algunas de ellas se han extendido a regiones geográficas muy amplias, ya que obtiene información y conclusiones sobre Estados Unidos, Europa, Canadá, Asia, África, etc., como puede verse en http: //www.gvu.gatech.edu/, y en la que también ponen de manifiesto en sus análisis no haber empleado un muestreo probabilístico.

A continuación se propone un método de muestreo probabilístico para la población de internautas P_I, el cual siempre será posible refinarlo según sea la información que se disponga en el marco de la encuesta.

En este caso de la Población Internet, el marco dará lugar a unas unidades primarias de muestreo que surgen de considerar $P_S = \{s_1, s_2 \ldots, s_M\}$ o conjuntos de "sites" de la Red que recubren la población P_I de internautas o e-población. Como unidades secundarias de muestreo se considerarán las visitas que realizan los internautas a los elementos de P_S, y que se representará por

$$P_v = \left\{P_{vs_1}, P_{vs_2}, \ldots, P_{vs_M}\right\} = \left\{\left\{v_{11}, v_{12}, \ldots, v_{1N_{v1}}\right\}, \ldots, \left\{v_{M1}, v_{M2}, \ldots, v_{MN_{vM}}\right\}\right\}$$

conjunto de las poblaciones visitas a cada "site" donde,

$$\sum_{j=1}^{M} N_{vj} = N_v$$

y las cúales verifican, desde el punto de vista del marco que se derivan, que las poblaciones de visitas tienen intersección vacía. Se seguirá utilizando para la selección de elementos la interceptación de éstos.

Tal como están definidas las unidades secundarias de muestreo, puede observarse la equivalencia que se da cuando un internauta se identifica en un "site" con el concepto de visitante que se ha planteado al estudiar la población audiencia - visitante.

5.1 Muestreo Probabilístico Bietápico

Sobre la población construida con las unidades primarias de muestreo, $P_S = \{s_1, s_2 \ldots, s_M\}$se aplicará un muestreo con probabilidades proporcionales al tamaño de una variable auxiliar X, que representa el número de visitas a los "sites" en un intervalo de tiempo de amplitud t fijado, siendo aconsejable que dicho intervalo sea lo más próximo posible al período de realización de la encuesta.

Por tanto si de P_S se extrae una muestra m_s de "sites" de tamaño q, la probabilidad de que el "site" s_i pertenezca a la muestra m_s, que se representará por $\pi_{si'}$ será

$$p_{s_i} = P[s_i \in m_s] = \frac{qx_i}{\sum_{i=1}^{M} x_i} \qquad i = 1,2,\ldots,M,$$

donde x_i será el número de visitas al "site" s_i, i=1,2,...,M, en un período de tiempo *t* fijado. Se supondrá que $qx_i < \sum_{i=1}^{M} x_i \quad \forall i$, si no fuera así, se seguirían métodos correctores como los propuestos en Kish (1965), Särndal, Swensson and Wretman (1992).

En la segunda etapa del muestreo, y para cada uno de los "sites" s_{ij}, j=1,2,...,q que forman la muestra de unidades primarias $m_s = \{s_{i_1}, s_{i_2}, \ldots, s_{i_q}\}$, se elegirán muestras independientes de visitas $m_{vsij} \subseteq P_{vs_{ij}}$, j =1,2,...,q, las cuales darán lugar a una muestra de visitas, m_V, formada por la unión de las muestras extraídas en cada "site", representada por la expresión conjuntista

$$m_v = \bigcup_{s_{i_j} \in m_s} m_{vs_{i_j}}$$

verificándose, por lo ya indicado, que $m_{vsij_r} \cap m_{vsij_s} = \varnothing \quad \forall i_{j_r} \neq i_{j_s}$.

Si el tamaño que se fije para la muestra de visitas m_v es n_v, el tamaño n_{i_j} para la muestra m_{vsij} extraída del "site" s_{ij} j =1,2,...,q , puede determinarse relacionándolos con alguna variable auxiliar. Ésta podrá ser el número de visitas que haya recibido cada "web site" en el período de tiempo *t* fijado previamente, tomándose en este caso

$$n_{i_j} = n_v \frac{x_{i_j}}{\sum_{j=1}^{q} x_{i_j}} \qquad j = 1,2,\ldots,q.$$

Como las muestras son extraídas de forma independiente en cada "site", podrá utilizarse uno de los métodos de muestreo probabilístico propuestos en el apartado de muestreo probabilístico para la población audiencia - visitas. La probabilidad de que una visita $v_k \in P_{vsi}$, para algún j =1,2,...,M y k =1,2,...,N_{vj}, pertenezca a la muestra m_v se expresará por

$$p_k = P[v_k \in m_v] = P[v_k \in m_{vs_i}, s_j \in m_s] = P[v_k \in m_{vs_i} \mid s_j \in m_s]\,P[s_j \in m_s] = \pi_{k/s_i}\pi_j,$$

y las probabilidades para que dos visitas $v_k, v_l \in P_v$ pertenezca a la muestra m_v será

$$p_{kl} = \begin{array}{lll} \pi_{kl/s_{ij}}\ \pi_{s_{ij}} & k \neq l \quad k,l = 1,2...N_v & v_k, v_l \in s_{i_j} \\ \pi_{k/si_{j1}}\pi_{k/s_{ij2}}\pi_{s_{ij1}s_{ij2}} & k \neq l \quad k,l = 1,2...N_v & v_k \in s_{i_{j1}}, v_l \in s_{i_{j2}} \end{array}$$

ya que las visitas estarán en la muestra extraída de un "site" o en dos muestras extraídas de distintos "sites", pues las visitas se identifican con un único "site". En la expresión anterior $\pi_{s_{ij1}\,s_{ij2}}$ es la probabilidad de elegir conjuntamente los "sites" $\left(s_{i_{j1}}, s_{i_{j2}}\right)$

Pero al estudiar la población Internet, las unidades de muestreo elementales con las que se tratará son los internautas precisados según los objetivos que se marquen al estudiar la población. Por tanto, se presenta una situación similar a la que se tenía previamente con los visitantes, únicamente que en este caso la multiplicidad ha de observarse a dos niveles, "multiplicidad dentro de cada "site'' y "multiplicidad entre "sites'', verificándose que $N \leq N_v$. Sobre la población $P_I = \{I_1, I_2, ...I_N\}$de internautas se definen las siguientes variables, para l =1,..,N;j =1,2,...,M:

$$\delta_{l_j}^{(1)} = \begin{Bmatrix} 1 & \text{si el internauta } I_l \text{ es identificado en el "site" } j-\text{ésimo} \\ 0 & \text{en otro caso} \end{Bmatrix}$$

de donde se tendrá que

$$\gamma_l^{(1)} = \sum_{j=1}^{M} \delta_{lj}^{(1)} \qquad l = 1,2,...,N$$

representa el número total de "sites" que el internauta l-ésimo, I_l, ha visitado.

Para cada "site" se define

$$\delta_{l\,jk}^{(2)} = \left\{ \begin{array}{c} 1 \text{ si el internauta } I_l \text{ es identificado con la visita} \\ k-\text{ésima}, v_k, \text{ al "site" } j-\text{ésimo} \qquad k = 1, 2..., N_{vj} \\ 0 \text{ en otro caso,} \end{array} \right\}$$

de donde

$$\gamma_{l\,j}^{(2)} = \sum_{j=1}^{N_{vj}} \delta_{ljk}^{(2)} = \#\, C_{lj}, \qquad l = 1, 2, \ldots, N \quad j = 1, 2, \ldots, M$$

siendo C_{lj} una clase de equivalencia tal como se definió en el caso de la población de visitantes:

$$C_{lj} = \left\{ v_{jk} \in P_{vsj} / v_{jk} \rightarrow I_l, k = 1, 2, \ldots, N_{vj} \right\} \quad l = 1, 2, \ldots, N, j = 1, 2, \ldots, M$$

Por tanto, el factor de ponderación a utilizar sobre los valores de la variable que se observen en el internauta l-ésimo, I_l viene dado por

$$\frac{\delta_{lj}^{(1)}}{\gamma_l^{(1)}} \frac{\delta^{(2)}{}_{ljk}}{\gamma_{lj}^{(2)}}; \quad k = 1, 2, \ldots, N_{vj}, \quad l = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, M$$

donde $\gamma_l^{(1)}$ y $\gamma_{l\,j}^{(2)}$ indican los grados o niveles de multiplicidad de los internautas y serán determinados por métodos como los ya indicados con anterioridad. Los estimadores a utilizar y sus propiedades pueden ser extraídas de los resultados recogidos en Sirken and Levy (1974), Lessler and Kalsbeek (1992), Bandyopadhayay and Adhikari (1993) y Byczkowski, Levy and Sweenwy (1998), entre otros.

5.2 Límite para el tiempo de realización de la encuesta

Al tratar con la población Internet se fija el tamaño muestral a obtener del dominio que constituye dicha población en la población de visitas, con el fin de que las estimaciones que se obtengan tengan la precisión adecuada. Esto hace que el tamaño de muestra a tomar en la población de visitas ha de ser mayor o igual al tamaño fijado para la muestra a tomar en la población Internet.

Sea $P_S = \{ s_1, s_2 ..., s_M \}$ la población de "sites" o unidades primarias de muestreo de la que se ha extraído una muestra de tamaño q, $m_s = \{ s_{i1}, s_{i2}, ..., s_{iq} \}$. De cada uno de los "sites" de m_s se extrae una muestra aleatoria de visitas independientes entre sí y, a su vez, independiente de las muestras extraídas del resto de los sites.

Para la muestra m_{vsij} extraída del site s_{i_j} se definen las variable indicadoras, para l = 1,2..,N, y j =1, 2,...,M

$$e_{ij} = \begin{cases} 1 \text{ si el internauta } l_l \text{ es observado en el "site" } j-\text{ésimo} \\ 0 \text{ en otro caso} \end{cases}$$

la cual se distribuye según una distribución Bernouilli con un parámetro que dependerá del "site" en que se encuentre el internauta

$$p_{ij} = P[e_{li} = 1] \qquad j = 1, 2, \ldots, q$$

Dichas variables son independientes. Para cada "site" se define, $X_{i_j} = \sum_{l=1}^{N} e_{li_j}$, que sigue una ley binomial $X_{ij} \in Bi\,(N, p_{ij})$ $\forall j=1,2,...,q$. Sea *n* la suma del número de internautas distintos observados en cada site, $n = \sum_{j=1}^{q} X_{i_j}$, y supongamos fijado n^*, número de internautas distintos en la muestra. El objetivo es estudiar la distribución de $n-n^*$, es decir, determinar k_0 tal que

$$1-\alpha = P\,[\, n-n^* \le k_0] .$$

Se tiene que

$$n - n^* = \sum_{1\le i<j\le q} N(i,j) - \sum_{1\le i<j<r\le q} N(i,j,r) + \cdots + (-1)^q N(1,2,\ldots,q)$$

donde

$$N(i_1, i_2, \ldots, i_j) = \sum_{l=1}^{N} e_{li_1}\, e_{li_2} \cdots e_{li_j}$$

representa el número de internautas que visitan los sites $s_{i1}, s_{i2}, \ldots, s_{ij}$ y sigue una distribución binomial, $Bi(N, \prod_{r=1}^{j} p_{i_r})$. Por el Teorema Central del Límite, cuando $N \to \infty$

$$\sqrt{N}\left(\frac{N(i_1,\ldots,i_j)}{N} - \prod_{r=1}^{j} p_{i_r} \right) \to N\left(0, \sigma^2(i_1,\ldots,i_j)\right)$$

donde $\sigma^{c}(i_1,i_2,\ldots,i_j)) = \prod_{r=1} p_{i_r} - \prod_{r=1} p_{i_r}^{c}$, y, por tanto,

$$\sqrt{N}\left(\frac{n-n^*}{N}-\mu\right)\to N(0,\sigma^2)$$

por lo que, k_0 puede aproximarse por

$$k_0=N\mu+\sqrt{N}\sigma z_{1-\alpha},$$

donde

$$\mu = E\left(\frac{n-n^*}{N}\right)=\sum_{1\le k<j\le q} p_i\, p_j - \sum_{1\le i<j<r\le q} p_i\, p_j\, p_r + \cdots + (-1)^q p_1\, p_2 \cdots p_q$$

$$\sigma^2 = \operatorname{var}\left(\frac{n-n^*}{N}\right)=\sum_{j=2}^{q}\sum_{1\le i_1<i_2<\cdots<i_j\le q}\sigma^2(i_1,i_2,\ldots,i_j)+\sum_{t=2}^{q}\sum_{r=2}^{q}\sum_{U}(-1)^{t+r}\sigma(i_1,i_2,\ldots,i_t;j_1,j_2,\ldots,j_r)$$

siendo

$$U=\{1\le i_1<i_2<\cdots<i_t\le q,\quad 1\le j_1<j_2<\cdots<j_r\le q;\quad (i_1,i_2,\ldots i_t)\ne(j_1,j_2,\ldots j_r)\}$$

$$\sigma(i_1,i_2,\ldots,i_t;j_1,j_2,\ldots,j_r)=\prod_{i\in I} p_i - \prod_{l=1}^{t} p_{i_l}\prod_{m=1}^{r} p_{j_m}$$

$$I=\{i_1,i_2,\ldots,i_t\}\cup\{j_1,j_2,\ldots,j_r\}$$

Expondremos a continuación dos casos particulares en cuanto a las probabilidades asociadas a los sites.

Caso 1: En el caso de que todos los sites lleven asociada la misma probabilidad, $p_i=a$, $i=1,\ldots,q$, expresiones de μ y σ^2 equivalentes a las obtenidas son

$$\mu(a;q) = (1-a)^q - 1 + qa$$
$$\sigma^2(a;q) = (1-a)^{q-2}a^2(1-a^2) + T_{q-2}$$

con

$$T_n=(n+1)a(1-a)-2a(1-a)^n-2(n-1)a(1-a)^{n+1}+(1-a)^{n+3}-(1-a)^{2n+2}$$

Caso 2: En este caso se consideran aquellas situaciones en las que existen dos conjuntos de sites con probabilidades de visita asociada a cada uno de ellos homogénea dentro de cada conjunto, $p_1 = p_2 = ... = p_k = a \neq b = p_{k+1} = p_{k+2} = ... = p_q$. Se obtienen las siguientes expresiones para μ y σ^2:

$$\mu(a,b;k,q-k) = (1-a)^k (1-b)^{q-k} - 1 + ka + (q-k)b$$
$$\sigma^2(a,b;k,q-k) = (1-b)^{q-k}\sigma^2(a;k) + R_{r-1}$$

con

$$R_n = ka(1-a) + (n+1)\,b\,(1-b) - ka\,(1-a)\,(1-b)^{n+1} - 2\,(n+1)\,(1-a)^k\,b\,(1-b)^{n+1} +$$

$$+ (1-a)^{2k}\,(1-b)\,n+1 - (1-a)^{2k}\,(1-b)^{2n+2}$$

Valores de μ, σ^2 y k_0, han sido obtenidos para α=0.05, N=3×10^6, N=4×10^6, y distintos valores de q,k,a,b, tal y como se muestran en la Tabla 2, concluyendo que el proceso de encuestación, para garantizar un número fijado de internautas distintos en la muestra, será finito y realizable.

5.3 Muestreo con Marcos Múltiples

Como ya se puso de manifiesto, es bastante improbable que los internautas que acceden a un "site" sean capaces de recubrir la población de internautas $P_I = \{I_1, I_2, ..., I_N\}$. Por tanto, se puede considerar que las unidades de muestreo que genera el marco de un "site" da lugar a un marco incompleto respecto a la población P_I. Esta situación da lugar a que aparezcan un conjunto de técnicas que traten de corregir dicha situación, como se observa en Kish (1965) y Lessler and Kalsbeek (1992). De las técnicas que se proponen, la más adecuada para aplicar en el caso de la población Internet, es el muestreo de marcos múltiples, que consiste en disponer de las unidades de muestreo procedentes de los marcos que proporcionan diferentes "sites", $P_s = \{s_1, s_2 ... s_M\}$, de forma que entre todos recubran la población P_I, y de cada uno de ellos se extrae una muestra mediante el muestreo probabilístico que se considere adecuado, siendo posible aplicar diferentes diseños muestrales en cada "site".

Para el tratamiento de las muestras que se obtienen desde los marcos múltiples, puede considerarse que los marcos que se generan desde cada site, dan lugar a un único marco con multiplicidad, y por tanto se aplicaría lo descrito previamente. No obstante, existe también la posibilidad de considerar todos los elementos y clasificarlos según presenten o no multiplicidades y agruparlos según el "site" en que hayan sido observados, y a continuación analizar cómo se ponderan las varia-

bles observadas para que los estimadores que se estudien tengan mínima varianza, como inicialmente propuso Hartley (1962) para dos marcos incompletos. También se utilizan otros criterios de optimalidad como pueden ser los relativos al costo.

Desde el trabajo de Hartley (1962), se han desarrollado estimadores y generalizaciones a más de dos marcos que permiten aplicar con garantías diseños muestrales con marcos múltiples, como se contempla en los trabajos de Bankier (1986), Skinner (1991), Kott, Amrhein and Hicks (1998) y Haines and Pollock (1998).

6. MUESTREO PROBABILÍSTICO EN UNA POBLACIÓN INTERNET ESPECIAL

En este tipo de población se sigue un procedimiento similar al anterior, es decir, se consideraría un marco donde las unidades primarias serán un conjunto de “sites” que recubrirán a la población Internet especial, y a continuación se tomarán como unidades secundarias las visitas a los respectivos “sites”, delimitadas por el tipo de especialización que se exijan a sus elementos, de lo que resulta necesario introducir en el cuestionario preguntas relativas a la especialización que se busca en la población objetivo, con el fin de discriminar los internautas que las verifican.

Por tanto, en este caso se abordará la encuesta de forma similar a lo propuesto anteriormente para la población de internautas.

7. MUESTREO PROBABILÍSTICO EN UNA POBLACIÓN NO SATURADA EN INTERNET

Cuando se disponga de una población no saturada en Internet, el marco de la población será, tal y como se indica en Cubiles et al.(2002), el correspondiente a dicha población, independientemente de que tenga o no acceso a la Red. El diseño muestral a aplicar también será independiente de los medios que dispongan los elementos para navegar por Internet.

No obstante, en el caso de disponer en el marco de información auxiliar que permitiera conocer la accesibilidad de los elementos a la Red, sería posible aplicar un diseño con un marco dual, y con ello aprovechar la disminución de costo que supone la realización de encuestas con entrevista en Internet. Si no fuera posible disponer de dicha información auxiliar para todos los elementos poblacionales, ésta se obtendría para los elementos que forman la muestra, y valorarse económicamente el dotar a los elementos muestrales del equipamiento informático necesario para responder la encuesta a través de la Red. Esto además supondría un incentivo para dichos elementos y, por consiguiente, un posible aumento en la tasa de respuesta de la encuesta.

En el caso de que la encuesta sea realizada a través del tiempo y se apliquen diseños muestrales como el rotativo o el muestreo panel, puede resultar absolutamente conveniente procurar que todos los elementos de la muestra tengan los medios necesarios para acceder a la Red con garantías para contestar la encuesta ya que ello redundará en la mejora de la tasa de respuesta, en la rapidez y fiabilidad de respuesta, al permitir una supervisión en tiempo real, en la mejora de la interactividad con los entrevistados y, seguramente, supondrá una disminución en el costo de la encuesta.

8. ERRORES

Como es conocido en toda encuesta por muestreo se producen diversos errores. En este caso se tendrán los errores propios del muestreo, que se generan al querer utilizar los procesos de inferencia y que han sido considerados previamente, ya que éstos pueden ser controlados por el tamaño de la muestra, y otros errores que surgen del resto de las acciones que lleva consigo la realización de una encuesta por muestreo probabilístico, los cuales pueden ser corregidos o amortiguados por técnicas como las que se aplican a encuestas realizadas por métodos de entrevista distinto a los de la Red. No obstante, en este caso además es conveniente tratar aspectos como los que se citan a continuación, por lo que puedan ayudar a disminuir los errores ajenos al muestreo.

Identificación: En los diseños que se han propuesto la identificación de los entrevistados resulta fundamental, porque ello nos sirve para corregir determinadas imperfecciones del marco, como eran las posibles multiplicidades, así como para la identificación de los visitantes o los internautas. El método de identificación a utilizar podría ser cualquiera aunque lo aconsejable es la identificación a través del correo electrónico, caso de que lo tenga. El hecho de aconsejar este método es porque en las encuestas a través de la Red pueden producirse circunstancias técnicas como caídas de sistemas, incompatibilidad de software, etc., que pueden provocar la no respuesta. La dirección electrónica puede servir para contactar con el entrevistado y así ofrecerle de nuevo la realización de la encuesta. Por ello, la identificación deberá ir próxima al inicio de la encuesta. No obstante, hay que indicar que en encuestas realizadas en Internet pedir la dirección electrónica supone cierto incremento en la tasa de no respuesta, como se indica en O'Neil and Penrod (2001).

Por otro lado hay que tener en cuenta la existencia de internautas con más de una dirección electrónica, por lo que se producirán multiplicidades si se utiliza más de una dirección en la identificación, resultando imposible detectarlo. Esta es una situación similar a cuando se utiliza el marco de las viviendas (hogares) o el método de entrevista telefónica donde los entrevistados puedan disponer de más de una

vivienda (hogar) o de más de un número de teléfono. En estos casos lo que suele suponerse es que el efecto en los errores de la encuesta es despreciable. Aquí se propone hacer la misma suposición, aunque en el caso de la referencia al correo electrónico pueda darse en segmentos importantes de la población de internautas un porcentaje de multiplicidad algo mayor que en los casos anteriores.

Por último, y sobre la identificación de los internautas, puede observarse la experiencia acumulada en las encuestas (no probabilísticas) realizadas en la Red, http: //www.gvu.gatech.edu/user-surveys/. En ellas pueden observarse desde porcentaje de internautas que dan direcciones falsas, o cómo los usuarios de Internet demandan conocer la utilidad que se le dará a la identificación que proporcionan. Todo esto puede ser útil para la estrategia a seguir al pedir la identificación de los entrevistados mediante la dirección electrónica.

Incentivos. En muchas de las encuestas realizadas en Internet suelen ofrecerse determinados incentivos a los entrevistados que cumplimentan la entrevista. Estos incentivos pueden ser de diferentes tipos, como la participación en sorteos donde se ofrecen diferentes premios a los ganadores, el incentivo de tipo personal o de tipo indirecto de aportación económica a alguna organización social. El efecto de los incentivos para incrementar la participación de los entrevistados es algo muy estudiado en encuestas donde se utilizan los métodos tradicionales de entrevista. Baste citar como trabajo representativo el de Church (1993). Respecto a la Red, ya se están realizando trabajos sobre este tema, obteniéndose hasta ahora conclusiones contradictorias por producirse unos en sentido positivo y otros en sentido negativo, como se reconoce en O'Neil and Penrod (2001).

Diseño del Cuestionario. Al ser la realización de entrevistas en Internet, un método de entrevista de los denominados autoadministrados, la presentación del cuestionario puede ayudar a incrementar la tasa de respuesta de los entrevistados, como muestran muchos trabajos de los realizados en este campo. Algunos de ellos han recogido o realizado experiencias para el caso de que las entrevistas se hagan en la Red, como es el de Bradley (1999), donde se indica que se da una mayor tasa de respuesta presentando el cuestionario de la encuesta cuando el entrevistado sea interceptado en el proceso de selección de la unidad de muestreo, en lugar de utilizar "banner" en diferentes páginas principales para invitar a la realización de la encuesta, o el de Dillman (2000), que configura una serie de etapas con recomendaciones precisas para construir el cuestionario de la encuesta, o el de Couper, Traugott and Lamias (2001), que experimentan con la forma de presentar las cuestiones en el cuestionario para incrementar la tasa de respuesta. Estos últimos reconocen la necesidad de realizar más experiencias en este campo, a pesar del desarrollo que ha tenido durante estos últimos años.

Marco Incompleto y Tasa de Recubrimiento. Como ya se indicó en el marco y en el método de muestreo a aplicar cuando la población objetivo sea representada por la población Internet o una especialización de ésta, los "sites" que constituyen las unidades primarias de muestreo pueden no recubrir la población de internautas de la población que se estudia. Como solución, ya se indicó la posibilidad de restringir la población objetivo a la población que recubrían las unidades primarias de muestreo. Si esa solución no se considerara la adecuada, debería estimarse el error que se está cometiendo al utilizar el marco incompleto. Esta estimación puede calcularse de muy diversas formas, pero quizás una de las más precisas, por lo que pueda suponer en rapidez y costo, es realizar una encuesta con entrevista telefónica, ya que en este caso el marco a utilizar en la encuesta siempre recubrirá de forma general a la población de interés.

El cuestionario de la entrevista telefónica tendría, como núcleo central del mismo, conocer si el entrevistado es internauta, y si lo fuera, conocer si durante el tiempo de realización de la encuesta accedió a la Red y al mismo tiempo accedió a al menos uno de los "sites" considerados como unidades primarias. Por tanto, una de las variables de interés para la subpoblación de elementos que entró en la Red durante el tiempo de realización de la encuesta es la siguiente:

$$\delta_i = \begin{cases} 0 \text{ si no accedió a ninguno de los "sites" de } P_S \\ \\ 1 \text{ si accedió al menos uno de los "sites" de } P_S \end{cases} \qquad i = 1, 2..., n_T$$

siendo $P_s = \{s_1, s_2, \ldots s_M\}$ y n_T el tamaño de la encuesta realizada mediante entrevista telefónica. Así, si

$$P_T = P\left[\delta_i = 1\right], \quad \forall i = 1,2,...,n_T,$$

entonces $1 - \hat{P}_T$ representa el error de recubrimiento que se tiene al utilizar como unidades primarias las correspondientes a P_S.

De idéntica forma podría plantearse esta medida para la población Internet en general, sin considerar los "sites" que son unidades primarias, en cuyo caso se definirá una variable indicador para la subpoblación de elementos que declaran ser internautas y que no han accedido a la red durante el tiempo de realización de la encuesta y los que si han accedido, lo que nos permitiría medir el error de recubrimiento.

A las variables indicadores antes citadas puede añadírseles los correspondientes niveles de especialización que se hayan fijado para la población objetivo, lo que nos mediría los correspondientes errores de no recubrimiento en dicha población.

9. CONCLUSIONES

La realización de encuestas por muestreo han ido incorporando a lo largo de la historia los avances tecnológicos en el campo de la comunicación.

La Red es un método de comunicación que se está imponiendo entre las personas y por tanto es útil para la realización de encuestas. De hecho, Internet ha sido utilizado prácticamente desde su creación. Sin embargo, resulta conveniente matizar que, por un lado, en estos momentos no puede utilizarse para hacer encuestas de población general debido al nivel de implantación actual, y, por otro lado, la metodología que se está aplicando para hacer encuestas desde la propia red no permite realizar inferencias estadísticas con el rigor que hay que exigirle a éstas. Por ello, se han propuesto diversos métodos para la realización de encuestas probabilísticas con el fin de tener alguna medida de los errores de muestreo que se producen al realizar las encuestas a través de Internet.

Los métodos propuestos dependen del tipo de población que se tenga como referencia para hacer la encuesta.

- Población Audiencia. Suponiendo que el proceso de entrevista es de interceptación, se consideran los siguientes casos:

– Audiencia-Visita: Se aplicará los métodos de muestreo probabilístico apropiados para una población general, dependiendo el proceso de selección de que se conozca o no el tamaño de la población.

– Audiencia-Visitante: En este caso se aplicarían las técnicas de muestreo apropiadas para una población en general, pero utilizando correcciones adecuadas para tratar un marco con multiplicidad. La encuesta audiencia-visitante es posible realizarla con un número de contactos aceptable.

- Población Precisada: En este caso las entrevistas no suponen un proceso de interceptación y, por tanto, antes de realizar la encuesta a través de la Red, será necesario contactar con los elementos de la muestra por el método que se considere adecuado. Asimismo se optará por algún diseño muestral probabilístico según los objetivos marcados en la encuesta.

- Población Internet. En este caso el muestreo probabilístico a aplicar puede plantearse de las siguientes dos formas:

– Muestreo probabilístico bietápico. En esta situación se consideran como unidades de muestreo de primera etapa los “sites” y unidades de segunda etapa los propios interanutas, pero utilizando correcciones adecuadas para las multiplicidades que se producen desde la población de internautas. Puede observarse (Tabla 2) que para los casos que se analizan no se necesitan un número elevado de interceptaciones de internautas para alcanzar el número de entrevistados fijados inicialmente.

– Muestreo con Marcos Múltiples. Aplicar este método sería adecuado si el número de sites que recubre la población de internautas es un número reducido.

• Población Internet Especializada. En este caso se utilizarán los mismos métodos de muestreo, introduciendo cierto método de screening para caracterizar la especialización. Se puede enmarcar el muestreo probabilístico dentro de los muestreos con más de una fase.

• Población no saturada en Internet. Dentro de este tipo de poblaciones puede incluirse la población general, y aunque no es adecuado en este caso hacer la encuesta con entrevistas a través de la Red, sí podría aprovecharse los elementos de la población con acceso a Internet y plantear un muestreo probabilístico con un marco dual. Incluso valorar qué significa económicamente para la encuesta que todos los elementos de la muestra tengan acceso a la Red.

Al esfuerzo aplicado al realizar métodos de muestreo probabilístico para controlar los errores propios del muestreo, se le han de unir aquellos otros que han de hacerse para amortiguar los errores ajenos al muestreo, como los que se han analizado previamente.

Tabla 1

COTAS SUPERIORES PARA EL NÚMERO DE VISITAS REQUERIDO, n_V, DADA LA PROBABILIDAD DE SER VISITANTE, p_V, Y EL NÚMERO DE VISITANTES, n_{V0}.

			α=0.10			
p_V \ n_{V_0}	100	200	400	500	1000	2000
1/4	444	863	1689	2099	4140	8198
1/3	331	644	1263	1570	3099	6141
1/2	218	426	837	1041	2058	4081
2/3	161	316	623	775	1535	3050
3/4	142	279	551	686	1361	2705
			α=0.05			
p_V \ n_{V_0}	100	200	400	500	1000	2000
1/4	458	882	1715	2129	4182	8256
1/3	341	658	1282	1591	3129	6181
1/2	224	434	847	1054	2074	4105
2/3	165	321	629	783	1546	3064
3/4	145	283	556	692	1369	2716
			α=0.01			
p_V \ n_{V_0}	100	200	400	500	1000	2000
1/4	486	920	1767	2186	4260	8366
1/3	361	684	1318	1631	3184	6259
1/2	235	449	868	1076	2106	4149
2/3	171	330	642	796	1565	3091
3/4	150	290	565	702	1383	2737

Tabla 2

VALORES DE $\mu(a,b;k,q-k)$, $\sigma^2(a,b;k,q-k)$, PARÁMETROS DE LA DISTRIBUCIÓN DE $n-n^*$, Y k_0 / $P(n-n^*=k_0)=1-\alpha$ PARA $\alpha=0.05$

a	*b*	*k*	*q*	μ	σ^2	$N=3x10^6$	$N=4x10^6$
						k_0	
0.01	0.0001	5	10	0.001015	0.000628	3115.388	4141.066
0.01	0.0001	5	25	0.001090	0.000707	3345.392	4446.988
0.01	0.0001	5	30	0.001115	0.000734	3423.464	4550.830
0.01	0.0001	5	100	0.001498	0.001137	4589.973	6102.807
0.01	0.0001	5	200	0.002124	0.001801	6494.316	8637.491
0.001	0.001	10	10	0.000045	0.000036	151.753	199.281
0.001	0.0001	5	10	0.000013	0.000009	46.113	59.991
0.001	0.0001	5	25	0.000022	0.000018	77.645	101.372
0.001	0.0001	5	30	0.000025	0.000022	89.570	117.064
0.001	0.0001	5	100	0.000102	0.000099	333.345	439.411
0.001	0.0001	5	200	0.000294	0.000295	931.820	1233.687
0.001	.0001	5	500	0.001454	0.001501	4471.749	5942.618
0.001	0.0001	10	25	0.000061	0.000052	203.144	267.181
0.001	0.0001	10	30	0.000067	0.000058	221.728	291.758
0.001	0.0001	10	100	0.000174	0.000167	558.863	738.572
0.001	0.0001	10	200	0.000411	0.000409	1289.688	1709.286
0.001	0.0001	10	500	0.001700	0.001755	5218.575	6936.780

REFERENCIAS

AZORÍN, F. Y SÁNCHEZ - CRESPO, J.L. (1986). «Métodos y Aplicaciones del Muestreo». Alianza Universidad Textos. Alianza Editorial.

BANDYOPADHAYAY, S. AND ADHIKARI, A.K. (1993). «Sampling from Imperfect Frames with Unknowm Amount of Duplications». *Survey Methodology,* 19. 193-197.

BANKIER, M.D. (1986). «Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys». *Journal of the American Statistical Association,* 81*,* 1074-1079*.*

BELLHOUSE, D.R. AND KULPERGER, R.J. (1991). «Computer Generated Simple Random Samples». *Communication in Statistics. Simulation and Computation,* 20, 539-550.

BISSELL, A.F. (1986). «Ordered Random Selection without Replacement». *Applied Statistics,* 35**,** 73-75.

BRADLEY, N. (1999). «Sampling for Internet Surveys. An examination of respondent for Internet research». *Journal of the Market Research Society,* 41, 387-395. BYCZKOWSKI, T.L., LEVY, M.S. AND SWEENEY, D.J. (1998). «Estimation in Sample Surveys Using Frames With a Many-to-Many Structure». *Survey Methodology,* 24, 21-30.

CHURCH, A.H. (1993). «Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta Analysis». *Public Opinion Quarterly,* 57, 80-91.

COUPER, M.P., TRAUGOTT, M.W. AND LAMIAS, M.J. (2001). «Web Survey Design and Administration». *Public Opinion Quaterly,* 65**,** 230-253.

CUBILES,M.D., MUÑOZ,M., MUÑOZ, J., PASCUAL, A. (2002).«e-Encuestas Probabilísticas I». Los Marcos. *En revisión.*

DILLMAN, D.A. (2000). Mail and Internet Surveys. «The Tailored Design Method». John Wiley & Sons, Inc.

FAN, C.T., MULLER, M.E. AND REZUCHA, I. (1962). «Development of Sampling Plans by Using Sequential (Item by Item) Selection Techniques and Digital Computers». *Journal of the American Statistical Association,* 57, 387-402.

HAINES, D.E. AND POLLOCK. (1998). «Combining Multiple Frames to Estimate Population Size and Totals». *Survey Methodology,* 24, 79-88.

HARTLEY. (1962). Multiple Frame Surveys. «*Proceedings of the Social Statistics» Section, American Statistical Association,* 203-206.

JOHNSON, N.L., KOTZ, S. AND WU, X. (1991). «Inspection Errors for Attributes in Quality Control». Chapman & Hall.

KENDALL, M.G. AND STUART, A. (1969). «The Advanced Theory of Statistics. Vol. 1». ed. Griffin, London.

KISH, L. (1965). «Survey Sampling». John Wiley & Sons.

KOTT, P.S., AMRHEIN, J.F. AND HICKS, S.D. (1998). «Sampling and Estimation from Multiples Frames». *Survey Methodology,* 24, 3-9.

LESSLER, J.T. AND KALSBEEK, W.D. (1992). «Nonsampling Error in Surveys». John Wiley & Sons, Inc.

MCLEOD, A.I. AND BELLHOUSE, D.R. (1983). «A Convenient Algorithm for Drawing a Simple Random Sample». *Applied Statistics,* 32**,** 182-184.

MORRIS, K.W. (1963). «A note on direct and inverse binomial sampling». *Biometrika,* 50**,** 544-545.

O'NEIL, K.M. AND PENROD, S.D. (2001). «Methodological variables in Web-based research that may affect results: Sample type, monetary incentives, and personal information». *Behavior Research Methods, Instruments, & computers,* 32**,** 226-233.

PINKHAM, R.S. (1987). «An Efficient Algorithm for Drawing a Simple Random Sample». *Applied Statistics,* 36, 370-372.

SÄRNDAL, C.-E.; SWENSSON, B. AND WRETMAN, J. (1992). «Model Assisted Survey» Sampling. Springer Verlag.

SIRKEN, M.G. AND LEVY, P.S. (1974). «Multiplicity Estimation of Proportions Based on Ratios of Random Variables». *Journal of the American Statistical Association,* 69, 68- 73.

SECO, A.; OLIMPIA, A. Y RAMOS, G. (2000). «Diccionario Abreviado del Español Actual». Ed. Aguilar.

SKINNER, C.J. (1991). «On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys». *Journal of the American Statistical Association,* 86, 779-784.

PROBABILISTIC E-SURVEYS II. PROBABILISTIC SAMPLING METHODS

SUMMARY

In this work there is approached fundamentally the study of the surveys that use the tool of Internet for its accomplishment. We centres on the exposition and development of probabilistic sampling designs that allow to realize surveys from the World Wide Web with the necessary accuracy to be able to infer the results obtained to the population under study, with certain reliability.

Keywords: survey, Internet, probabilistic sampling design.

AMS Classification: 62D05, 62P99