ESTADÍSTICA ESPAÑOLA
Vol. 44, Núm. 151, 2002, págs. 281 a 305

e-Encuestas Probabilísticas I. Los Marcos

por
MARÍA DOLORES CUBILES DE LA VEGA
Departamento de Estadística e Investigación Operativa
Universidad de Sevilla
MARÍA MACARENA MUÑOZ CONDE
Departamento de Estadística e Investigación Operativa
Universidad de Sevilla
JUAN M. MUÑOZ PICHARDO
Departamento de Estadística e Investigación Operativa
Universidad de Sevilla
ANTONIO PASCUAL ACOSTA
Departamento de Estadística e Investigación Operativa
Universidad de Sevilla
Centro Andaluz de Prospectiva

RESUMEN

Se formalizan los conceptos que surgen al plantear la realización de encuestas entre los usuarios de Internet. Para ello se define el concepto de e-encuesta probabilística, que extiende la definición de encuesta probabilística al ámbito de Internet. Tras discutir algunos problemas que surgen al intentar identificar de forma unívoca a los usuarios de la Red, se describen las características más importantes de la población internauta española. Según el grado de disponibilidad de acceso a Internet, se establece una clasificación de las poblacio-

nes objetivo de las e-encuestas probabilísticas en: Poblaciones Saturadas en Internet y Poblaciones No-Saturadas en Internet.

Palabras claves: encuestas, Internet, marco, audiencia.

Clasificación AMS: 62D05,62P99

1. INTRODUCCIÓN

El espectacular desarrollo experimentado por Internet en los últimos años ha posibilitado el acceso a información de todo tipo y volumen por parte de los usuarios de esta "Red de Redes", si bien la calidad de esta información depende considerablemente de los portales visitados, por lo que su uso, además de un gran número de ventajas, también puede presentar algunos inconvenientes. En cualquier caso, Internet puede ser un recurso muy valioso para la profesión estadística, por cuanto constituye una fuente de datos de campos muy diversos, y que abre nuevos retos en cuanto a la adaptación y desarrollo de las técnicas estadísticas a utilizar dentro del acceso a los datos de Internet y su análisis.

En este trabajo nos centramos en uno de esos problemas, concretamente el de la realización de encuestas entre los usuarios de Internet. Para ello, en el siguiente apartado, tras un resumen histórico de las encuestas y una descripción de los procesos de encuestación que actualmente se realizan en Internet, se define el concepto de e-encuesta probabilística, que permite acomodar el concepto de encuesta probabilística al ámbito de Internet. En el apartado tercero se describe la población internauta española que puede ser objetivo de las e-encuestas. El cuarto apartado presenta una discusión de la problemática que surge al intentar caracterizar el marco de una e-encuesta.

2. e-ENCUESTAS PROBABILÍSTICAS

El vocablo **encuesta** tiene como principales sinónimos, voces como: "averiguación", "pesquisa", "información", "indagación", "investigación", "estudio y "sondeo", y si se recurre a Seco, A.; Olimpia, A.; y Ramos, G. (2000), se define **encuesta**, como *"consulta hecha a numerosas personas, para conocer determinadas circunstancias políticas, sociales o económicas, o el estado de opinión sobre un tema".*

El hombre ha tenido siempre un gran afán por "averiguar", "informarse" e "investigar" sobre todo lo que le rodeaba, por lo que el empleo de encuestas aparece con

el hombre y por tanto se remonta al principio de los tiempos. En particular, ya en torno al año 3.000 antes de Cristo se encuentran referencias sobre Censos, en los que el término "numerosas personas" se identifica con "toda la población"[1]. Si además se consideran las necesidades que tenía el hombre de conocer lo que le rodeaba cuantificándolo, lo que fue previo incluso a la aparición del número según indica Barrow (1997), se pueden situar los inicios de los procesos de obtención de información en el año 35.000 antes de Cristo.

La evolución de la Humanidad ha permitido también la evolución de los métodos y técnicas utilizados por el hombre para la realización de sus actividades, y como no puede ser menos, estos avances se han producido también en el campo de la obtención de información mediante encuestas, aunque en este caso no es hasta el siglo pasado cuando se comienza a tener una teoría con fundamento científico para la realización de encuestas.

En este sentido, se pasó de estudiar la población en su totalidad (censos) a estudiar el máximo número posible de elementos de ella, siguiendo cierta sistematización en la captación de la información y con la única restricción de que los elementos pertenezcan a la población. Esto puede observarse en trabajos como el de 1.860 de Henry Mayhew titulado *London Labour and the London Poor* o el de Charles Booth de 1.890, que lleva por título *The Life and Labour of the People of London*, en los que se comenzaba a utilizar ya el proceso de entrevistas en la realización de encuestas, hecho que, tal como se indica en Moon (1999), puede considerarse como un primer paso en dirección a lo que hoy en día se conoce como encuesta por muestreo.

Sin embargo estos métodos de obtención de información, donde a la muestra o conjunto de elementos de la población que proporcionan información para la encuesta, sólo se le exige que el número de elementos a entrevistar sea lo mayor posible, sin llegar a cubrir toda la población, presentan la imposibilidad de generalizar los resultados que se obtengan a la población de la que proceden por los errores que puedan cometerse. Sobre este hecho existen muchas evidencias empíricas, baste citar como uno de los ejemplos más llamativo del error de proyección de la muestra a la población, la encuesta realizada por The Literary Digest en 1936 sobre las elecciones presidenciales en Estados Unidos tal como se recoge en Bradburn y Sudman (1988).

Los primeros intentos por desarrollar un método que permitiera generalizar los resultados que se obtengan en una encuesta a la población de la que proceden los entrevistados, se deben a Anders Nicolai Kiaer, que en el Congreso del Internatio-

1 En lo que sigue, se asumirá como población un conjunto de elementos

nal Statistical Institute celebrado en Berna en 1.895, presentó un método que imponía ciertas condiciones o restricciones a la participación y/o selección de los elementos en las encuestas que denominó “Muestreo Representativo”, aunque hay que decir que este método no fue aceptado de forma inmediata. Tuvieron que transcurrir varias decenas de años para que la comunidad científica lo reconociera como un método estadístico capaz de realizar inferencias con ciertos niveles de precisión sobre la población en la que se realizaba la encuesta, lo que ocurrió en el Congreso del International Statistical Institute celebrado en Roma en 1925, donde, entre la terminología que se introdujo apareció el término “muestreo aleatorio”, que significaba que “todos los elementos de la población tenían la misma probabilidad de ser elegidos”, tal como se recoge en Kruskal y Mosteller (1980). A partir de ese año 1925 se produce un gran desarrollo en lo que se conoce como teoría de muestreo.

Se consideran encuestas por muestreo, aquellas en las que la selección de los elementos de la población a proporcionar información para la encuesta se produce mediante ciertos criterios o condiciones prefijadas de forma no subjetiva. Si estos criterios se fundamentan en las probabilidades de selección de los elementos o de la muestra, se tendrán las encuestas por muestreo probabilístico, que permiten generalizar o proyectar las conclusiones que se obtengan en la encuesta a la población, acompañando las proyecciones de ciertas medidas de la precisión que tiene el proceso de inferencia o de generalización. El conjunto de elementos de la población que constituyen la muestra en la aplicación de un muestreo probabilístico recibe el nombre de muestra probabilística.

De forma paralela al desarrollo de la teoría que sustenta el muestreo probabilístico, se fueron incorporando distintos métodos para la obtención de la información en las encuestas, los cuales se clasifican de forma genérica en: *métodos administrados*, que se caracterizan por presentar ciertos niveles de interacción entre el elemento de la muestra o entrevistado y el entrevistador o persona que demanda de forma directa la información al entrevistado, y *métodos autoadministrados*, en los que no existen intermediarios entre el entrevistado y el cuestionario de la encuesta, aunque Díaz de Rada (2000) distingue en este último caso el “entrevistador virtual” y la “ausencia total de interacción”. Existen numerosos trabajos sobre la efectividad y la aplicabilidad de muchos de los métodos de entrevistas que se engloban en las clases genéricas citadas anteriormente, comprobándose en algunos de ellos cómo se ha ido incorporando el desarrollo tecnológico a la recogida de información en las encuestas, baste citar a Groves, R.M. (1989), Bosch y Torrente (1993), Díaz de Rada (2000), Dillman (2000), etc.

En estos momentos, se está incorporando Internet, o la Red, a través de todas sus componentes, ya sea la World Wide Web, el correo electrónico, la transferencia

de ficheros (FTP), gopher, usenet newsgroups, etc, al campo de las encuestas. Esto está dando lugar a la necesidad de formalizar nuevos conceptos, actualizar y adaptar técnicas ya existentes a estas nuevas tecnologías, desarrollar nuevo *software* que responda a las demandas que se impongan desde el punto de vista de los métodos estadísticos, etc.. De estos esfuerzos cabe esperar grandes ventajas, ya que Internet aportará a las encuestas diversas mejoras, como es en lo relativo a la interacción con los entrevistados, la posibilidad de abordar de forma cómoda poblaciones en ámbitos geográficos más amplios, la reducción de costos en las encuestas, la disminución del número de errores en la cumplimentación de los cuestionarios, la rapidez en la realización, etc.

Dependiendo del recurso de la Red que se utilice para la realización de encuestas tendrán que considerarse unos aspectos u otros de las encuestas por muestreo. Así, en el caso del correo electrónico, se tiene que algunas de sus condiciones y problemas son muy similares a los que se presentan en las clásicas encuestas por correo. Sin embargo no ocurre lo mismo con el uso de otras herramientas, como puede ser la utilización de la World Wide Web (www, Web, w^3), donde resulta necesario introducir nuevos conceptos y realizar modificaciones de determinados métodos estadísticos para poder realizar encuestas por muestreo probabilístico, como se recoge a continuación.

Hace ya algún tiempo que se están realizando encuestas en la Web. Visitando ciertas páginas de determinados “web site” o “site”, se tiene la posibilidad de responder a determinadas preguntas que constituyen el cuestionario de una encuesta. En algunas “páginas web” se le ofrece a sus visitas un número importante de encuestas sobre temas muy diversos invitándolos a participar en aquellas que crean más convenientes. Los métodos que se siguen en estas encuestas sólo persiguen obtener un número de respuestas lo mayor posible sin una mayor exigencia y por tanto no están sometidas al rigor científico necesario para poder trasladar los resultados que se obtengan a una población definida con antelación a la obtención de la muestra. Esto es reconocido por algunos “sites”, que hacen encuestas como las indicadas y que en sus páginas recogen frases como las que siguen:

“Este sondeo de opinión es una encuesta voluntaria para nuestros usuarios y no puede proyectarse con bases científicas para aplicarse a cualquier otro grupo de población. Ofrecemos estas encuestas para dar a nuestros usuarios la oportunidad de compartir sus opiniones sobre temas concretos”, en la dirección electrónica http://vr.harrispollonline.com/voting.

“Esta encuesta no es científica, responde tan sólo a las respuestas voluntarias de los lectores que desean exponer su opinión”, en http://www.elpais.es/encuestas.

En estos momentos puede decirse que la mayoría de encuestas que se realizan en Internet carecen del rigor necesario, lo que impide realizar proyecciones, y cuando se realizan encuestas bajo muestreo probabilístico, se hacen sobre poblaciones muy controladas o precisadas, o se utilizan métodos de entrevista mixtos, o es una población muy reducida, etc. Se concluye por ello la necesidad de abordar diseños muestrales probabilísticos que permitan realizar encuestas desde la Web con capacidad de generalizar los resultados que se obtengan a la población o poblaciones que se consideren en cada momento, y esto tiene su importancia ya que a las ventajas de carácter general, indicadas previamente para la realización de encuestas en la Red, puede unírsele las relativas a aplicaciones como las de interés informativo y de valoración de noticias de utilidad a los medios de comunicación. También puede citarse el interés comercial, por todo lo que se relaciona con el e-comercio, con el e-marketing y con el desarrollo de nuevos servicios en la misma Red, o encuestas de cualquier tipo como las relativas a la salud o incluso en el campo de la Estadística Pública. En este sentido se propone como definición de **e-encuesta probabilística** la siguiente:

Definición. Se denomina e-encuesta probabilística a toda encuesta donde el proceso de entrevista a los elementos de la muestra se realiza en la Red a través de la www y en la que se conoce la probabilidad de selección de los elementos y/o de la muestra.

Una primera aproximación al diseño y realización de e-encuestas probabilísticas se conseguirá adaptando a Internet las etapas y métodos que se siguen en el diseño general de una encuesta por muestreo probabilístico.

3. POBLACIÓN, INTERNET Y CARACTERIZACIÓN DEMOGRÁFICA

Al pretender utilizar Internet como medio para la realización de encuestas, resulta necesario caracterizar la población que podría participar en las e-encuestas probabilísticas y a la que se denominará e-población, que estará constituida por aquellas personas que pueden acceder a los medios adecuados para la cumplimentación del cuestionario de la encuesta a través de Internet. Dado que éstos son los recursos necesarios para tener la conexión adecuada a Internet, puede decirse que la e-población estará formada por los usuarios de Internet o población de internautas.

Tener acceso a la Red, no supone estar censado por ningún tipo de organización pública o privada, por ello el conocimiento que se tiene de la población antes indicada y su caracterización por sus principales variables demográficas se ha conseguido mediante la realización de encuestas, utilizando muchos de ellos como

medio para hacer las entrevistas, el correo, el teléfono o bien la “entrevista cara a cara”

En España una de las principales referencias para caracterizar a los usuarios de Internet es el Estudio General de Medios (EGM), cuyos principales resultados pueden encontrarse en http://www.aui.es. Se destacan a continuación algunos de ellos, que pueden resultar de interés para el diseño de encuestas donde la entrevista se realice a través de la Red.

La población de referencia que toma el EGM es la población española de 14 años o más, siendo 34.818.000 el número de elementos que considera para dicha población, y por tanto la referencia para realizar proyecciones de población desde el EGM.

La evolución del porcentaje de personas que utilizan ordenador y los usuarios de Internet en los últimos años se representan en la figura 1, de la que puede concluirse una cierta tendencia hacia la convergencia de los usuarios de ordenadores y los de Internet.

Sin embargo los usuarios de Internet no recubren la población española de 14 o más años, siendo la tasa de recubrimiento de la población en Febrero/Marzo de 2001 igual al 19,8%. Algo similar con más o menos recubrimiento ocurre en otros países como se desprende de estudios realizados por Computer Industry Almanac y Nielsen/NetRatings Inc. a los que también se puede acceder desde las páginas de http://www.aui.es.

Del EGM se observa que los usuarios de Internet o de la Red acceden a ella desde diferentes lugares. Si se considera la multiplicidad media de acceso a Internet, definida como el cociente entre el número total de usuarios de Internet que surgen de considerar los diferentes lugares de accesos (casa, trabajo, universidad, otros) y el número total de usuarios de Internet, se tiene para Febrero/Marzo de 2001, una multiplicidad media de acceso de 1,233, como se desprende de los datos que generan la figura 2.

La asiduidad en el acceso no es uniforme en la población, teniendo los usuarios conductas muy dispares. No obstante éstas tienden a confluir a comportamientos de mayor asiduidad en la conexión a la Red, como se puede ver en la representación gráfica (figura 3) de los datos del EGM.

Si se consideran determinadas variables demográficas para desagregar la población como son el sexo y la edad, se contempla en la figura 4 que la distribución de sexos de los usuarios de Internet de más de 14 años se aleja de la distribución de sexos para dicha edad en la población española, donde las mujeres representan aproximadamente el 51,5% de la población (I.E.A, 2001). No obstante, la tendencia

es la de aproximarse a dichos valores, algo que está ocurriendo en otros países con una mayor tradición en el uso de la Red que la existente en el caso español.

En la figura 5 se observa una distribución de usuarios de Internet según los distintos grupos de edad muy diferente de la distribución existente en la población española. En particular, la población más joven se encuentra sobrerepresentada como usuaria de Internet, según se desprende de datos como los siguientes: en la encuesta Feb/Mar 2001, se tiene que aproximadamente el 4% de los usuarios de Internet tienen una edad superior a los 54 años, mientras que en la población española dicho grupo de edad representa el 30,17%; entre 25 y 34 años se encuentra el 32,3% de los usuarios de Internet, frente a un 19% en la población total. Habrá que esperar cierto tiempo para que dichos porcentajes se aproximen, aunque lo ideal para la realización de e-encuestas es que se produjera la universalización en el acceso a la Red de los distintos grupos de edad, lo que supondría la universalización de cualquier población objetivo.

En la figura 6 se representa la evolución de la e-población según clases sociales observándose una gran diferencia en el uso de Internet entre las clases sociales Alta, Media-Alta y Media-Media, y el resto.

Un análisis sociológico más pormenorizado del EGM puede encontrarse en Wert (2000). Estudios similares al EGM hay hechos en otros países, en particular Couper (2000) contiene muchas referencias para Estados Unidos. Muchos de estos estudios tienen unas peculiaridades y obtienen unas conclusiones parecidas a las que se han indicado previamente.

Por último hay que indicar que no todas las encuestas que se realizan sobre usuarios de Internet, permiten realizar generalizaciones sobre el comportamiento y la diversidad socio-demográfica de la e-población, ya que los entrevistados suelen ser usuarios voluntarios conseguidos a través de anuncios (“banners”) colocados en páginas web estratégicas u otros procedimientos similares, y con ello no cumplen una de las condiciones básicas de toda encuesta realizada por muestreo probabilístico.

4. EL MARCO Y LOS USUARIOS DE INTERNET

Del marco de una encuesta se han dado numerosas definiciones, a continuación se recoge como referencia la dada en Lessler y Kalsbeek (1992).

“El marco consta de los materiales, procedimientos y dispositivos que identifican, distinguen, y permiten acceder a los elementos de la población objetivo. El marco está compuesto de un conjunto finito de unidades (unidades de muestreo) a las que se aplica el esquema del muestreo probabilístico. Las reglas o mecanismos

para enlazar las unidades del marco a los elementos de la población son una parte fundamental del marco. El marco también incluye información auxiliar (medida de tamaño, información demográfica) utilizada para (1) técnicas de muestreo especiales, tales como estratificación y selección de la muestra con probabilidad proporcional al tamaño, o (2) técnicas de estimación especial, tal como la estimación de la razón o regresión".

Enlazar los elementos de la población con unidades de muestreo que sean lo más identificables posibles condiciona a veces la forma de realizar las entrevistas "cara a cara", "por correo" y "por teléfono", y recíprocamente, la forma de realizar las entrevistas pueden delimitar el marco a utilizar. En este sentido y para muchas poblaciones, suelen tomarse como unidades de muestreo, los hogares (viviendas) para los dos primeros métodos de entrevista citados y para el caso de la encuesta telefónica se considera el número de teléfono asociado al hogar (vivienda), bien por las referencias de las guías telefónicas existentes o por el recorrido de las posibles números de teléfonos, como se propone en Waksberg (1978). En todos estos casos también ocurre que el entrevistador puede acceder a los entrevistados que forman la muestra cuando lo desee (cuestión aparte es que la vivienda esté vacía o no se coja el teléfono, etc.) y además en ningún caso supone costo económico alguno para el entrevistado.

En el caso de la población usuaria de la Red, podía pensarse en enlazar los elementos de la población con la dirección IP ("Internet Protocol") del ordenador que utiliza para acceder a la Red, esta direccción identifica al ordenador conectado a Internet y es proporcionada por el "Network Information Center" como se recoge en Castells (2001). Sin embargo, no existe una relación biunívoca entre los ordenadores con acceso a la Red y su dirección IP, ya que existe lo que se conoce como direcciones "IP dinámicas", que son las direcciones que proporcionan las ISP ("Internet Service Provider"), es decir, las empresas u organizaciones que ofrecen el servicio de conexión a Internet. Cuando un cliente de una ISP accede a la red, recibe una dirección IP, la cual queda libre en cuanto el usuario se desconecta, por lo que dicha dirección IP puede ser asignada a un nuevo cliente que inicie su conexión. Este proceso se describe más ampliamente en Foo, S., Hui, S.Ch., Yip, S. y He, Y. (1997). Además la asignación de las direcciones IP tiende a realizarse por amplias zonas geográficas, lo que dificulta a veces la localización de los usuarios de Internet.

Por consiguiente, considerar como unidades de muestreo los ordenadores, o más precisamente sus direcciones IP, no cumple las condiciones básicas para la identificación de los elementos de la población, como se exige en la construcción del marco en la realización de cualquier encuesta por muestreo. A ésto se ha de unir el hecho de que la entrevista por Internet puede suponer cierto "costo" para el

entrevistado, salvo condiciones especiales, y acceder al entrevistado no suele producirse por la voluntad expresa del entrevistador.

Por tanto, las unidades de muestreo que formarán el marco de una encuesta donde la entrevista se realizará a través de Internet, y sobre las que se calcularán las probabilidades a utilizar en el muestreo probabilístico, serán los algunas veces llamados "surfistas de la Red" pero con una referencia al intervalo de tiempo de realización de la encuesta. Esta referencia temporal permite caracterizar el número de unidades de muestreo que forman el marco, permitiendo amortiguar los errores que puede provocar en el muestreo probabilístico el continuo crecimiento de la población de internautas y la movilidad o volatilidad existente en las conductas de los usuarios de Internet. Además, las unidades de muestreo del marco habrán de ser precisadas aún más cuando se especifique la población objetivo.

5. POBLACIÓN OBJETIVO Y MARCO

Como se ha visto previamente la e-población es un subrecubrimiento de la población general de muchos países, lo que ocurre por el hecho de que ser usuario de Internet no es una cualidad universal, ni siquiera una cualidad extendida entre los elementos de muchas poblaciones, por lo que en estos momentos la población de internautas no recubrirá a muchas de las poblaciones objetivo que se planteen para realizar encuestas por muestreo. En esta línea puede adaptarse lo recogido en Trewin y Lee (1988) para las encuestas telefónicas, indicando que cuando la implantación de Internet en una población objetivo sea del orden del 80% o superior, puede utilizarse la entrevista a través de Internet como un método para la realización de encuestas en dicha población. Esta población objetivo será calificada como "*Población Saturada en Internet*", y también puede definirse como aquella población en la que el 80% o más de sus elementos constituyen una e-población. En caso contrario, si la e-población representa un porcentaje inferior al 80% de la población, a ésta se le denominará, "*Población No - Saturada en Internet*" y la e-población se tomará como una población generadora de un marco dual para la población objetivo o como medio para controlar ciertos niveles de calidad de la encuesta. Esto hace que se proponga descomponer las poblaciones objetivo sobre las que se realizan encuestas con entrevista a través de la Red en dos grandes bloques con sus correspondientes subclasificaciones, como se indica a continuación:

Poblaciones Saturadas en Internet:

1. Población Audiencia

2. Población Precisada

3. Población Internet

4. Población Internet Especial

Poblaciones No - Saturadas en Internet

Estos tipos de poblaciones deben considerarse modelos genéricos de población objetivo a los que podrán ajustarse muchas de las encuestas que se realicen en Internet y sobre ellos puede admitirse incluso que son poblaciones sin límites geográficos, siendo ésta otra de las grandes ventajas que las entrevistas en la Red aportan a los métodos de encuestación.

5.1. Poblaciones Saturadas en Internet

5.1.1. Población Audiencia y Marco

En este caso la población objetivo se precisará al fijar un "site" o una red de "web sites" de la Red (en adelante se hará referencia a un único "site") y considerando sobre él su audiencia. Para precisarla aún más puede adaptarse lo citado en Callejo (2001): *la población audiencia* puede definirse como el conjunto de usuarios de Internet que acceden a la "web site" fijada, por los motivos que sean, durante un período de tiempo dado.

La población audiencia de un "site" no está formada por un número estable o fijo de usuarios de Internet, baste considerar que siempre es posible hablar de "audiencia potencial", que incluiría a todos los usuarios de Internet que tienen los medios para acceder al "site" fijado y por tanto puede afirmarse que la población audiencia siempre será una población abierta, término que hay que considerar en un doble sentido, por un lado por lo que significa de audiencia potencial, y por otro por el hecho de que la población de usuarios de la Red está en continuo crecimiento.

Por tanto para la población audiencia del "web site" en estudio, y por lo indicado de forma genérica sobre el marco, se considerarán como unidades de muestreo los usuarios de Internet que acceden al "web site" durante el período de tiempo de realización de la encuesta. Estas unidades se precisarán más según los aspectos particulares que se fijen en los objetivos de la encuesta.

Las unidades de muestreo de la población audiencia pueden ser consideradas en general bajo dos objetivos diferentes, según que se considere la Población Audiencia de los Visitantes o bien la Población Audiencia de Visitas. En el primer caso surge en el marco el problema de la multiplicidad de sus elementos, lo que procede de las posibles repeticiones en el acceso al "site", lo que no ocurre en el caso de considerar la Población Audiencia de Visitas. Por ello el marco de la población audiencia conviene analizarlo según los dos esquemas mencionados.

A) VISITANTES Y POBLACIÓN AUDIENCIA

Si se consideran como unidades de muestreo los elementos de la e-población que acceden al "site", fijando como objetivo estudiar los visitantes, el marco presenta multiplicidades entre sus elementos, debido a la repetición o recurrencia con que los visitantes entren en el "site" bajo estudio. Para corregir las multiplicidades del marco, pueden seguirse métodos como los indicados en Kish (1965) y en Lessler y Kalsbeek (1992) entre otros, si bien el método mas fiable y por tanto recomendado es el de la supresión de las multiplicidades. Para ello en las e-encuestas probabilísticas puede pensarse en las conocidas "cookies", que ya se utilizan en otras actividades que se realizan en la Red, incluso en el mismo campo de las encuestas, como indican Wang, Dziuban y Hartman (2000).

Las "cookies" son pequeños ficheros en formato texto que el "web site" puede implantar en el sistema informático del elemento de la población audiencia que accede a él. Para ello, el servidor Web transmite una cabecera de texto ("la cookie") dentro del flujo de información en formato HTML utilizado en las transmisiones bajo protocolo HTTP. Esta cabecera de texto incluye información sobre el usuario, por ejemplo sus preferencias según la información accedida en el servidor Web. La "cookie" es recibida por el navegador del usuario (en general Netscape o Internet Explorer), y se almacena en el disco duro del usuario dentro de un fichero especial que contiene la lista de "cookies" (por ejemplo, Netscape lo nombra "cookies.txt"). Cada vez que el usuario acceda posteriormente a dicho servidor Web, el navegador envía la "cookie" de forma automática al servidor Web. Todo este proceso suele hacerse sin el conocimiento del usuario, y uno de sus objetivos es identificar de forma unívoca a los sistemas que acceden al "site". Sin embargo, el trabajo a desarrollar y el conocimiento proporcionado por parte de las "cookies" sobre los ordenadores donde estén instaladas puede ser muy diverso, como se recoge en trabajos como los de Highland (1997) y Peng y Cisna (2000), entre otros. En algunos casos el conocimiento extraído conduce a cuestionar en cierta medida la legalidad de las "cookies", al ser capaces de proporcionar información que puede ser considerada "reservada" por el elemento de la e-población.

Otro problema que presentan las “cookies” es que si bien pueden identificar de forma precisa los ordenadores desde donde se accede al “site”, no son capaces de identificar a los internautas visitantes. Como se sabe por encuestas realizadas a los usuarios de Internet, éstos pueden acceder desde distintos sistemas informáticos o al contrario, un mismo sistema informático puede ser utilizado por más de un internauta. Por tanto, las “cookies” no resuelven el problema de multiplicidad en el marco, e incluso pueden conducir a un subrecubrimiento de la población audiencia dado que existen grupos de usuarios compartiendo el mismo ordenador. A todo lo indicado se ha de unir la posibilidad que tienen los ordenadores de protegerse de las “cookies”, mediante la configuración adecuada de las aplicaciones utilizadas para acceder a Internet.

En resumen, la propuesta que se hace por parte de algunos autores de utilizar las cookies para resolver ciertos problemas de identificación, no es muy adecuada para resolver los problemas de multiplicidad del marco, a lo que hay que unir las cuestiones legales que pueden plantear su uso.

Por ello la solución para resolver las posibles repeticiones en el marco es identificar a los visitantes con la correspondiente dirección IP asociada al ordenador a pesar de lo indicado sobre ella, y pedirles que se identifiquen como un usuario, siendo lo recomendable que ésto se produzca mediante la dirección electrónica y una contraseña (“password”). En algunos casos se puede completar el proceso de identificación solicitando del entrevistado algunos datos particulares, como fecha y lugar de nacimiento, o el domicilio, y que serán confrontadas con los datos reales que la organización que realiza la encuesta puede poseer previamente en sus bases de datos. El trabajo ya citado de Wang, Dziuban y Hartman (2000) recoge una experiencia de este tipo con una población formada por los alumnos de una universidad.

De lo analizado puede concluirse que la Población de Visitantes es considerada como un dominio o subpoblación de la e-población que visita al “site”, sobre la que se aplica un método de discriminación fundamentado en la identificación de los visitantes, tal como se ha indicado.

B) VISITAS Y POBLACIÓN AUDIENCIA

Encuestar la población de visitas al “site” puede tener también su interés, ya que entre los objetivos de la encuesta puede incluirse el estudio de las características de los elementos de la población que más reinciden en su acceso al “site”, el acceso a través de las diferentes páginas del “site”, la procedencia desde otra “web site”, la variabilidad en la información o servicio que demanda etc.

Cuando la población objetivo es la de las visitas al "site", no se plantea el problema de multiplicidad. No obstante, es recomendable seguir un proceso de identificación como el que se realiza con la población de los visitantes, lo que permite medir la reincidencia de las visitas de una forma precisa, y además permite la posibilidad de analizar la evolución en las respuestas de los internautas u ofrecerle al entrevistado el ratificarse o no en la encuesta realizada con anterioridad, e incluso proponerle nuevos cuestionarios según el número de veces que vaya accediendo al "web site".

Tanto para el marco de la población de visitantes como de las visitas, se requiere información auxiliar: la organización y estructura de información del "web site", número de visitas y visitantes que se ha producido en un determinado tiempo, y todas aquellas características que se conozcan de la audiencia, procurando la máxima desagregación e indicando su variabilidad en el tiempo.

5.1.2. Población Precisada y su Marco

Se considera como población identificada a aquella población saturada en Internet, de la que es posible disponer o conseguir de una forma cómoda un listado con todos sus elementos conteniendo la información apropiada para su localización. Dicho listado generará las unidades de muestreo del marco, al que habrá que unir todas la información auxiliar de que pueda disponerse para diseñar el método de muestreo a aplicar de la forma más precisa posible. Poblaciones precisadas pueden ser, los miembros de asociaciones científicas, alumnos universitarios, ciertos clientes de empresas, las propias empresas, investigadores, etc.

En este tipo de poblaciones es conveniente analizar y estimar el nivel de subrecubrimiento que se presenta al utilizar la e-población, por la utilidad que ello tiene en el análisis y conclusiones de la encuesta.

5.1.3. Población Internet y su Marco

En esta población objetivo se incluyen todos los usuarios de Internet independientemente del "site" que visiten. Ésta es una e-población con un recubrimiento total de la población objetivo que se considere. No obstante los objetivos de la encuesta pueden provocar la necesidad de precisar ciertas características del internauta entrevistable, como se muestra a continuación.

Un elemento de la e-población o internauta, es cualquier persona que navega por Internet, como se recoge en muchos diccionarios dedicados a la Red, pero esta definición así dada abarca a personas con comportamientos muy dispares respecto a su conexión y navegación por la Red tal como desprende de muchas encuestas y estudios, como el EGM ya citado con antelación. Por ello, cualquier definición de

elemento de la e-población necesite que se precisen los comportamientos a considerar, según los objetivos de la encuesta. Así, Callejo (2001), define los elementos de la e-población como: *"todos aquellos individuos que declaran haber accedido a Internet al menos una vez en los últimos siete días, sea cual fuese el propósito de ese acceso a la red, el uso, el lugar, la vía, etc."*.

Esta última definición solo depende la asiduidad en el acceso a la Red. Variando la asiduidad en Internet y los grados de independencia o dependencia de factores como los citados previamente (el uso, el lugar, la vía, etc.), se obtienen diferentes definiciones para los elementos de la Población Internet. Cualquiera que sea la definición que se adopte para la población Internet y sus elementos, éstos siempre serán entrevistados por un procedimiento de interceptación en la Red, pero sin fijar un "site" o red de "sites", como ocurre en la población audiencia.

Por consiguiente, en el marco se considerarán dos tipos de unidades de muestreo: las unidades primarias de muestreo, que estarán constituidas por el conjunto de "sites" que recubren la población Internet objeto de la encuesta, y las unidades secundarias, que serán los internautas de la población cuando acceden a alguno de los "sites" que son unidades primarias, todo ello referenciado al tiempo de realización de la encuesta.

Independientemente de la definición que se adopte para la población Internet, ésta siempre se fundamenta en internautas - visitantes, por lo que el marco constituido por las unidades de muestreo antes descrito presenta el problema de multiplicidad ya descrito y debido a la repetición de visitas a un mismo "site", pero que en este caso incluye también la multiplicidad que se deriva de acceder los visitantes a distintos "sites". Al margen de las posibles soluciones que se adopten desde un punto de vista analítico para resolver el problema de multiplicidad, siempre será conveniente que los internautas se identifiquen de la forma descrita con anterioridad.

En la representación de la población de internautas mediante las unidades de muestreo previamente descrita, se plantean también ciertos problemas de subrecubrimiento, ya que el conjunto de los "sites" que se consideren pueden no recubrir la población Internet objeto de la encuesta. Ésto puede resolverse redefiniendo la población como la población de internautas de los "sites" tomados como unidades primarias de muestreo, en caso contrario será necesario medir el efecto del subrecubrimiento en las estimaciones o proyecciones que se realicen.

También es necesario resaltar que el tiempo de realización de la encuesta puede producir efectos de subrecubrimiento en la población de Internet, por el hecho de que algunos internautas no accedan a la Red en dicho período. En este caso, el problema de subrecubrimiento puede amortiguarse exigiendo ciertos niveles de

asiduidad en la navegación del internauta, o adoptando un tiempo de realización de la encuesta amplio. En este último caso debe valorarse el efecto que dicho tiempo puede tener en la composición de la población de internautas.

A este marco de la población Internet será necesario añadir toda la información auxiliar que puedan ofrecer "los sites", que son unidades primarias de muestreo y que puede ir desde sus contenidos hasta las características que conozcan de su audiencia y de la información que ofrezcan. Una cuestión que resultará imprescindible para los diseños muestrales que se planteen es el número de visitas a las unidades primarias.

5.1.4. Población Internet Especial y su Marco

Ésta será una población saturada en Internet, donde los elementos de la e-población habrán de ser estudiados o analizados con el fin de conocer si verifican una determinada especialización exigida para formar parte de la población objetivo, la cual no tienen por qué cumplirla única y exclusivamente los elementos de la e-población. Los elementos que verifiquen la condición exigida al definir la población constituirán lo que se denomina población Internet especial.

En este caso, al ser la población que se trata una población saturada, se trabajará con la e-población y por tanto se empleará el marco que se ha definido para la población Internet, para a continuación seguir con sus elementos un proceso de identificación que permita caracterizar la subpoblación que da lugar a la población Internet especial, según lo que se prevea en los objetivos de la encuesta. Al igual que en todos los casos anteriores, la población Internet y el marco correspondiente tendrán la referencia temporal al período de realización de la encuesta.

En este tipo de población saturada, se plantean desde un punto de vista práctico dos problemas de subrecubrimiento: el primero puede deberse a las unidades primarias de muestreo, "los sites", tal como se indicó en el caso anterior, y el segundo puede deberse al propio recubrimiento de la e-población sobre la población objetivo del estudio, y por tanto puede resultar conveniente estimar las correspondientes tasas de recubrimiento.

5.2. Poblaciones No - Saturadas en Internet y su Marco

Estas poblaciones se caracterizan por el hecho de no poder efectuar en ella encuestas donde el proceso de entrevistas se realice basándose sólo en la Red, ya que como se indicó la e-población tiene un recubrimiento inferior al 80% de la población. Dentro de estas poblaciones no - saturadas en Internet se encuentran prácticamente todas aquellas para las cuales se pretenda realizar una encuesta sociodemográfica, según se vio al analizar dichas variables en la e-población.

Por tanto en las poblaciones no - saturadas en Internet se construirá el marco adecuado a los objetivos que se tengan para la encuesta, independientemente de la e-población que se disponga y como ya se dijo esta última población puede ser utilizada en un planteamiento de marco dual o como medio para controlar la calidad de la encuesta.

No obstante, puede resultar conveniente en muchas ocasiones comparar el costo que supone realizar la encuesta con entrevistas realizadas directamente, ya sea cara a cara o telefónicamente, con el que supondría el hecho de dotar a los elementos de la muestra que no pertenezcan a la población de internautas de los medios necesarios para que lo sean. También es necesario resaltar en este último caso lo que ello puede significar en aminorar los errores ajenos al muestreo.

REFERENCIAS

BARROW, J.D. (1997). «Por qué el Mundo es Matemático». *Grijalbo Mondadori*.

BOSCH,J.LL.C. Y TORRENTE, D. (1993). «Encuestas telefónicas y por Correo. Cuadernos Metodológicos». *Centro de Investigaciones Sociológicas*.

BRADBURN, N.M. AND SUDMAN, S. (1988). «Polls and Surveys, Understanding what They Tell Us». *Jossey - Bass Publishers.*

CALLEJO, J. (2001). «Investigar las Audiencias. Un análisis cualitativo. Papeles de Comunicación 34». *Paidós*.

CASTELLS, A. (2001). «Diccionario de Internet. Todos los términos usados en la WWW*». Ediciones Deusto.*

COUPER, M.P. (2000). «Web Surveys. A Review of Issues and Approaches». *Public Opinion Quarterly, 64, 464-494*.

DIAZ DE RADA, V. (2000). «Utilización de Nuevas Tecnologías para el Proceso de "Recogida de Datos" en la Investigación Social mediante» *Encuesta. Reis, 91, 137-166.*

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

FOO, S., HUI, S.CH., YIP, S.W. AND HE, Y. (1997). «Approaches for Resolving Dynamic IP Addressing». *Internet Research: Electronic Networking Applications and Policy*, *7, 208 - 216*.

GROVES, R.M. (1983). «Survey Errors and Survey Costs». *John Wiley & Sons*.

HIGHLAND, H.J. (1997). «The Threats on The Web». *Computers & Security*, *16, 365-368.*

I.E.A. (2001). «Anuario Estadístico de Andalucía 2001». *Instituto de Estadística de Andalucía. Junta de Andalucía.*

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

KRUSKAL, W. AND MOSTELLER, F. (1980). «Representative Sampling, IV: the History of the Concept in Statistics, 1895-1939». *International Statistical Review*, *48*, *169-195*.

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

MOON, N. (1999). «Opinion Polls. History, Theory and Practice». *Manchester University Press*.

PENG, W AND CISNA, J. (2000). «HTTP Cookies-a Promising Technology». *Online Information Review*, *24, 150-153*.

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

TREWIN, D. AND LEE, G. (1988). «International Comparisons of Telephone Coverage. Telephone Survey Methodology»*. Ed. R.M. Groves, Bemer, P.P., Lyberg, L.E., Massey, J.T., Nichols II, W.L. and Waksberg, J. John Wiley & Sons*.

WAKSBERG, J. (1978). «Sampling Methods for Random Digit Dialing»*. Journal of the American Statistical Association*, 73 , 40-46.

WANG, M.C., DZIUBAN, C.D., HARTMAN, J.L. (2000). «A Web-Based Survey System for Distributed Learning Impact Evaluation with SAS System»*.* «Proceedings of the Twenty-Fifth» *Annual SAS Users Group International.Conference (accesible electrónicamente: www2.sas.com/proceedings/sugi25/25/ad/25p024.pdf).*

Figura 1

EVOLUCIÓN DE LOS DATOS

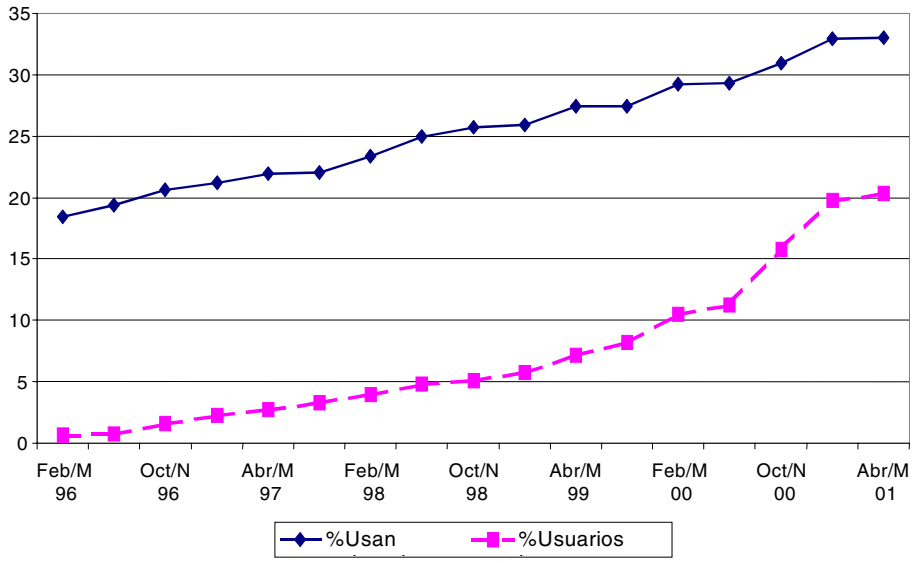


Figura 2

LUGAR DE ACCESO A INTERNET

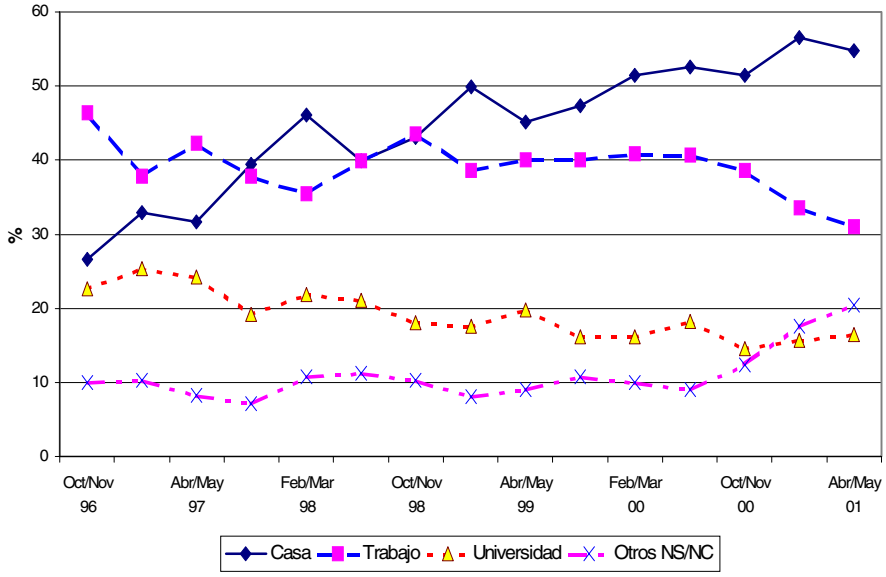


Figura 3

FECHA DEL ÚLTIMO ACCESO A INTERNET

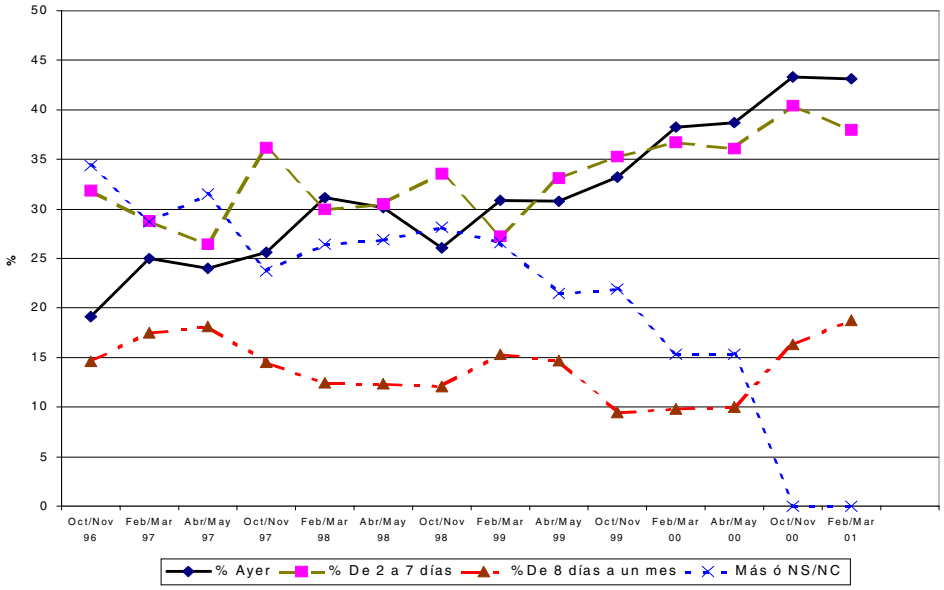


Figura 4

PERFIL DE USUARIOS DE INTERNET SEGÚN SEXO

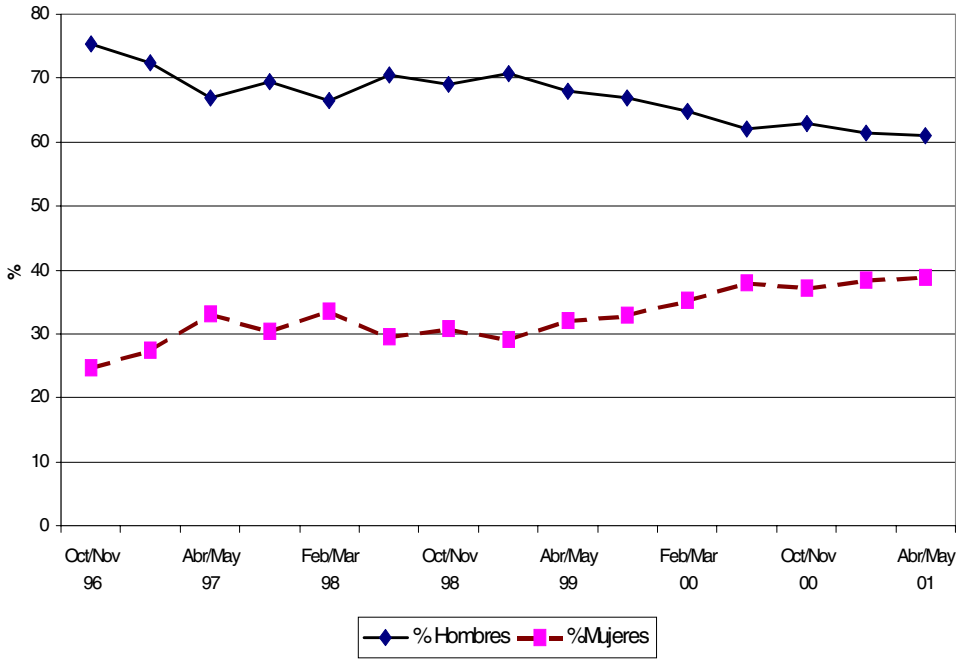


Figura 5

PERFIL DE USUARIOS DE INTERNET SEGÚN EDAD

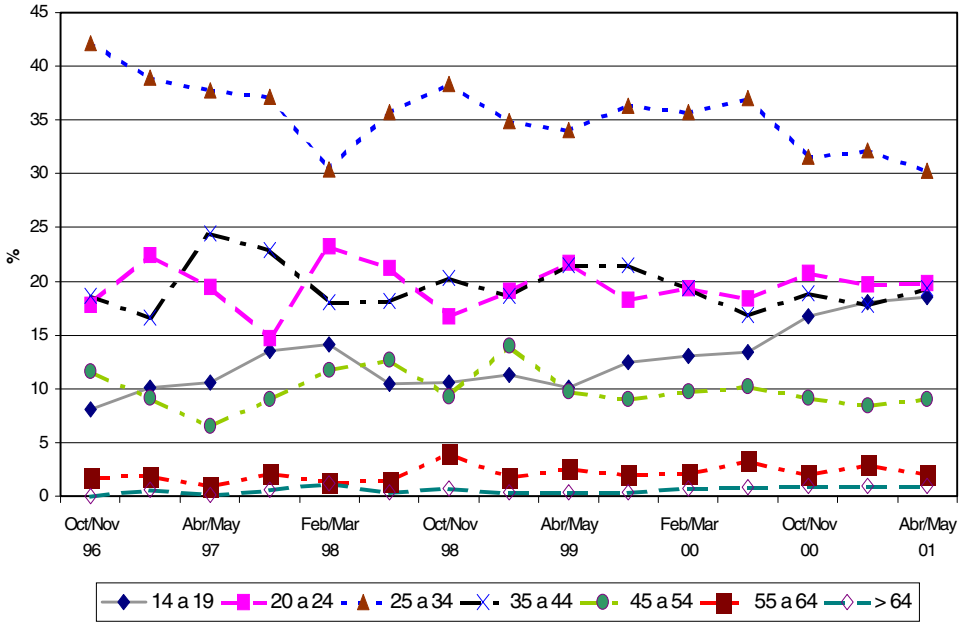
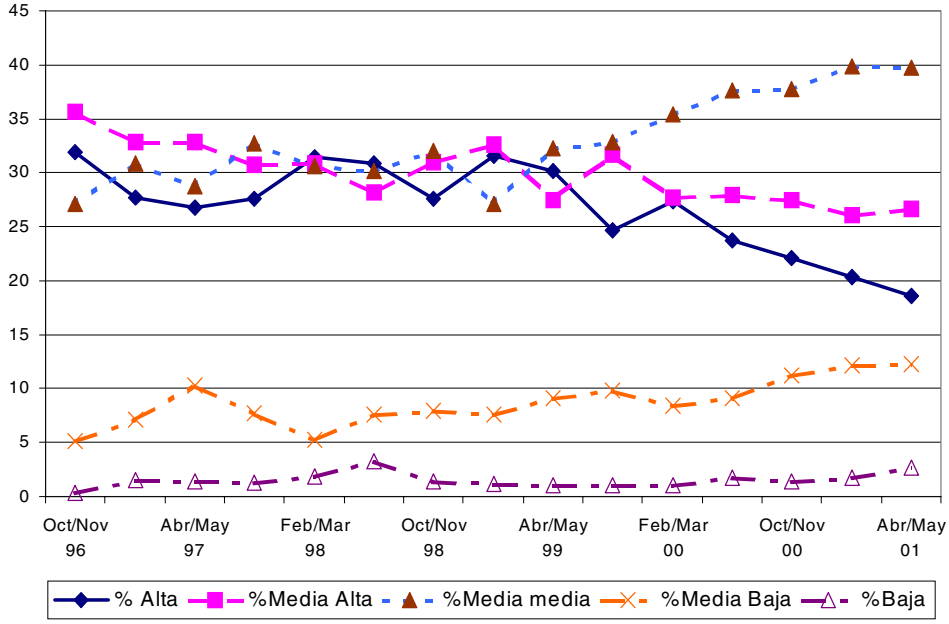


Figura 6

USUARIOS DE INTERNET SEGÚN CLASE SOCIAL



PROBABILISTIC E-SURVEYS I. THE FRAMES

SUMMARY

The topics arising from Internet surveys are formalized in this work. So we define the notion of probabilistic e-survey, extending the probabilistic survey definition to the Internet field. We discuss several problems when we try to unanimously identify the Internet users, and we describe the most important characteristics of the Spanish Internet-navigate population. Considering the level of availability in Internet access, we establish a classification of target populations of the probabilistic e-surveys: Internet Saturated Populations and Internet Not-Saturated Populations.

Key words: surveys, Internet, frame, audience.

AMS Classification: 62D05,62P99