Introducción a los métodos de encuestación y muestreo estadístico

1. Resumen teórico
    1.1. Muestreo: definición y utilidades
    1.2. Conceptos básicos
    1.3. Tipos de error
    1.4. Tipos de muestreo: Probabilístico, intencional, por cuotas.
    1.5. Diseño del muestreo
    1.6. Formas de administración: Personal, correo, telefónica, internet
2. Métodos de muestreo
    2.1. Muestreo aleatorio simple
    2.2. Muestreo estratificado
    2.3. Muestreo por conglomerados
3. Caso práctico: muestreo en población general.
    3.1. Tipo de muestreo: probabilístico o por cuotas.
    3.2. El seccionado.
    3.3. Estratificación
    3.4. Determinación del tamaño de la muestra.
    3.5. Elección aleatoria de los conglomerados.
    3.6. Definición de las cuotas.

1. RESUMEN TEÓRICO

El propósito de las encuestas es obtener información de las poblaciones. Existen dos estrategias posibles de recolección de datos:

- Examinar todas las unidades de la población: censo
- Examinar, de acuerdo a unas pautas preestablecidas, un subconjunto de la población.

La recogida de datos es un punto importante en todo estudio estadístico, ya que el mejor método de análisis resulta poco provechoso si se aplica sobre datos inválidos.

1.1. Muestreo: definición y utilidades

Si pretendemos conocer una serie de características de un cierto colectivo y éste posee un elevado número de elementos su estudio exhaustivo se hace difícil. Para solucionarlo se extrae un subconjunto de los elementos originales y, mediante la información suministrada por dicha selección, se obtendrán conclusiones de las características de la población completa.

La *teoría del muestreo* tiene como objetivo suministrar la metodología que guíe los problemas de recogida de información. Aunque la práctica nunca es igual a los modelos teóricos, estos son indispensables como guía para establecer las condiciones adecuadas en la elección de las unidades últimas a encuestar. El muestreo es la herramienta para seleccionar la parte de la población cuya observación permitirá extender la información obtenida al conjunto de la población objetivo del estudio. Para que las conclusiones sobre la población sean adecuadas es necesario que la selección de las unidades se realice de tal manera que las unidades escogidas sean lo más representativa posible de la población total y, para esto, es fundamental planificar adecuadamente el método usado para la selección.

1.2. Conceptos básicos

Se denomina ***población*** a un conjunto de unidades de la cual se desea obtener información. Estas unidades pueden se personas, viviendas, familias, escuelas,…, y la información a obtener puede ser de cualquier tipo, relación con la actividad, ingresos por familia…

El estudio que se desea realizar sobre la población, se manifiesta través de una serie de características, desconocidas de antemano, que pueden ser de muy diversa naturaleza. Por ejemplo, en la población de alumnos matriculados en una universidad, una característica posible de estudio podría ser el número de cursos que cada uno lleva estudiando, al menos matriculado, en la misma, contando con el actual. Esta sería una característica intrínsecamente cuantitativa numérica. Un alumno puede llevar matriculado un curso, dos

cursos, ocho cursos, etc. Otra característica podría ser si disfruta no de algún tipo de beca durante el presente curso. Esta sería una característica de tipo cualitativo, SÍ o NO, intrínsicamente no numérica.

Una característica cualquiera bajo estudio será representada matemáticamente por una variable, *Y*, que llamaremos ***variable de estudio***. Los valores que toma dicha variable sobre las unidades poblacionales son desconocidos de antemano. Para denotarlos emplearemos la siguiente notación,

$$\boldsymbol{X = (x_1, x_2, \ldots, x_N)}$$

siendo y_i el valor de la variable de estudio para la unidad poblacional i. Así, para el ejemplo anterior, denotando por el número de cursos cada alumno lleva matriculado, si $x_{567}=4$, esto quiere decir que el alumno número 567 lleva matriculado en la universidad 4 cursos, incluyendo el presente.

En la población es posible medir en cada unidad a estudiar una o varias características. A partir de estos resultados se pueden llegar a calcular valores como el consumo medio por familia, la proporción de miembros del hogar en paro…, a estos valores se les conoce como ***parámetros o características poblacionales.*** Se trata de un valor numérico que describe una característica de una población. Los parámetros son valores numéricos constantes, es decir, no son variables. Definida una población cualquiera y un parámetro en ella, ese parámetro sólo puede tomar un valor numérico concreto. Habitualmente los parámetros de interés serán la media y los porcentajes.

Para elegir las unidades que van a formar la muestra es necesario disponer de un conjunto real de unidades que se ajuste lo mejor posible al conjunto que forma la población objetivo. A la lista de estas unidades a partir de las cuales es posible seleccionar la muestra la denominaremos ***marco***.

Muestra: una muestra es un subconjunto de elementos de una población. Para extraer conclusiones validas e imparciales referidas a todos los elementos de la población a partir de la observación de sólo unos pocos elementos, es necesario, que la muestra utilizada sea representativa de la población; esto se consigue mediante las “técnicas de muestreo”.

Tamaño muestral: es el número de elementos que constituyen la muestra. Los elementos que componen la muestra se seleccionarán de la población generalmente de forma aleatoria, por tanto una muestra de tamaño “n” puede interpretarse como una variable aleatoria n-dimensional cuya distribución de probabilidad dependerá de la distribución de probabilidad F(X) de la población y del tamaño muestral “n”.

Llamamos ***espacio muestral*** al conjunto de todas las muestras posibles extraídas por un procedimiento de muestreo. Al procedimiento mediante el cual se extrae la muestra se denomina ***muestreo***.

Estadístico: un estadístico es un valor numérico que describe una característica de una muestra. Su valor concreto depende de los valores de la muestra seleccionada en la que es calculado. Es evidente que de una población cualquiera es posible extraer más de una muestra diferente del mismo tamaño, por tanto el valor de un estadístico varía de una muestra a otra. Un estadístico no es un valor numérico constante (como lo es un parámetro), sino que es una variable: su valor concreto depende de la muestra en la que es calculado.

Algunos de los estadísticos principales son: la media muestral, la varianza muestral, el total muestral y la cuasivarianza muestral, la proporción muestral, el máximo y mínimo de la muestra. Un estadístico que se utiliza para estimar un parámetro desconocido de la población recibe el nombre de estimador.

Existen propiedades deseables para un estimador, la insesgadez y la mínima varianza. Un estimador es *insesgado* del parámetro que desea estimar si su esperanza matemática (media) coincide con el parámetro poblacional. Además de la insesgadez del parámetro es deseable que el estimador tenga una varianza mínima. ¿Qué conseguimos con estas propiedades de insesgadez y mínima varianza? Conseguimos un estimador preciso, es decir, que los posibles valores que pueda tomar el estimador estén próximos al parámetro con una elevada probabilidad.

Algunos estimadores fundamentales:

Estimador de la esperanza matemática: Consideremos las muestras de tamaño n, de una característica sobre una población que viene expresado a través de una v.a. X con media y varianza:

$$X_1, X_2, \ldots, X_n, \qquad \begin{cases} \mathbf{E}[X_i] = \mu \\ \mathbf{Var}[X_i] = \sigma^2 \end{cases}$$

El estimador *media muestral* que denotaremos normalmente como $\overline{X}$ (en lugar de μ) es:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$ y verifica:

$$\boxed{\mathbf{E}[\overline{X}] = \mu} \qquad \boxed{\mathbf{Var}[\overline{X}] = \frac{\sigma^2}{n}}$$

Por tanto es un estimador insesgado.

Estimador de la varianza: A la hora de elegir un estimador de $\sigma^2 = \mathrm{Var}[X]$, podemos comenzar con el estimador más natural:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

El valor esperado del estimador no es σ^2, y por tanto el estimador para la varianza no es insesgado. Más aún, su esperanza vale

$$\mathrm{E}[S^2] = \frac{n-1}{n}\sigma^2.$$

Para tener un estimador insesgado de la varianza introducimos la cuasivarianza muestral $\hat{S}^2$ que se define como

$$\hat{S}^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Es inmediato comprobar que realmente este estimador es insesgado

$$\mathrm{E}\left[\hat{S}^2\right] = \mathrm{E}\left[\frac{n}{n-1}S^2\right] = \frac{n}{n-1}\mathrm{E}[S^2] = \frac{n}{n-1}\,\frac{n-1}{n}\sigma^2 = \sigma^2.$$

1.3. Tipos de error

Cualquier encuesta por muestreo tiene asociados una serie de errores que se pueden clasificar en dos tipos: errores debidos al muestreo y errores ajenos al muestreo.

1.3.1. Errores debidos al muestreo

Al extraer la muestra, los datos obtenidos a partir de la muestra nos permiten inferir unos valores aproximados de la población total. A estos valores se les denomina ***estimaciones*** estas estimaciones llevan unido un error, el ***error debido al muestreo***. Cuanto menor sea este error mayor es la precisión de las estimaciones.

Existe un procedimiento que permite obtener la muestra estableciendo a priori el error que estamos dispuestos a tolerar, este hecho nos permitirá afirmar con certeza que las

afirmaciones que realicemos sobre los valores de la población no conllevarán un error de muestreo superior al límite establecido a priori.

Este procedimiento consiste, en primer lugar, en elegir el estimador de forma que tenga la mayor precisión posible, es decir, que se comporte de la misma forma que el parámetro poblacional. Para esto es deseable que cumpla las propiedades de insesgadez y mínima varianza.

Para controlar la precisión de un estimador hay que estudiar su *error cuadrático medio*. Así, dado un parámetro poblacional, que denotaremos por θ, buscamos un estimador θ~. Se define el error cuadrático medio de un estimador como:

$$ECM(\tilde{\theta}) = var(\tilde{\theta}) + (sesgo\ (\tilde{\theta}))^2 = var(\tilde{\theta}) + (E(\tilde{\theta}) - \theta)^2$$

A partir del ECM vamos a definir dos tipos de errores, el error de muestreo y el *error de muestreo relativo*:

- Se define el error de muestreo del estimador θ~ como la raíz cuadrada de su ECM
- Se define el error de muestreo relativo como el cociente entre el error de muestreo y la esperanza del estimador.

Existen procedimientos estadísticos que nos permiten establecer a priori el error de muestreo que estaríamos dispuestos a tolerar y elegir la muestra estableciendo este límite. Esto nos llevaría a poder afirmar que las conclusiones que realicemos sobre la población conllevarían un alto grado de certeza.

1.3.2. Errores ajenos al muestreo

Además de los errores de muestreo existen otros tipos de error, los errores ajenos al muestreo. Estos errores no se asocian al proceso de muestreo. Se dividen en dos grupos:

- Errores de observación: se deben a la recogida, registro o procesamiento incorrecto de los datos, se subdividen en errores de sobrecobertura (el marco contiene elementos que no pertenecen a la población objetivo), de medida (diferencias entre el valor observado y el valor verdadero) y de procesamiento (se trata de errores que se producen en el tratamiento de los datos una vez recogidos).
- Errores de no observación: Se producen cuando no es posible obtener información deseada para ciertos elementos de la población o cuando no es posible incluir elemento de la población en la muestra. Se trata de errores de cobertura y errores de falta de respuesta.

1.4. Tipos de muestreo: Probabilístico, intencional, por cuotas.

Existen tres tipos básicos de muestreo: muestreo probabilístico, muestreo intencional u opinático y muestreo por cuotas.

- *Muestreo probabilístico*: Diremos que un muestreo es probabilístico cuando se conoce a priori la probabilidad de obtener cada una de las muestras que es posible seleccionar. Por tanto, cada una de las muestras tiene asociada una probabilidad de extracción, la suma de todas las probabilidades es igual a la unidad. Si las muestras se escogen según un modelo aleatorio, se puede modelizar el comportamiento de los estimadores y cuantificar el error cometido en las estimaciones.
- *Muestreo no probabilístico*: Es aquél en que la elección de las unidades muestrales no se realiza de forma aleatoria siguiendo la Teoría de muestreo o no se conoce la probabilidad de selección de cada una de las posibles muestras. Hay básicamente dos tipos: muestreo intencional u opinático y muestreo por cuotas.

Hay ocasiones en que la selección de las unidades no es aleatoria, sino que la persona que realiza la selección procura encontrar esa representatividad, aunque esta depende de su intención u opinión. Este tipo de muestreo, llamado *muestreo intencional u opinático*, aplicado bajo condiciones correctas, puede dar resultados útiles, pero carece de base teórica y, por tanto, imposibilita el cálculo de su error.

Otro tipo de muestreo muy utilizado en la práctica es el *muestreo por cuotas*: en una primera etapa se descompone la población en grupos de elementos (excluyentes y exhaustivos) con características comunes definidas previamente. Cada grupo definido representará un determinado porcentaje en la población, dicho porcentaje es a lo que se acostumbra a denominar cuota. La segunda etapa consiste en elegir una muestra que refleje exactamente esa proporción. Normalmente, las cuotas más utilizadas son el sexo y la edad.

Hay que señalar que el muestreo probabilístico es el más adecuado siempre que sea posible utilizarlo, ya que sobre él hay una teoría científica que permite elevar los resultados obtenidos para la muestra al conjunto de la población. Además, en este tipo de muestreo se pueden calcular errores de estimación asociados a las estimaciones.

1.5. Diseño de una encuesta por muestreo

El diseño de una encuesta es un aspecto fundamental a tener en cuenta, ya que el fallo en uno de los aspectos que comprende puede invalidar la encuesta. Como fase previa a su

planeamiento se deben fijar cuáles son sus objetivos, qué información se necesita para cumplirlos y de qué medios se dispone.

De forma esquemática el diseño de una encuesta consta de los siguientes aspectos:

1. Trabajos preliminares: determinación del objetivo u objetivos de la encuesta, descripción de la población a estudiar, definiciones, modelo esquemático de tablas y cuestionario.

Se debe comenzar por establecer las necesidades a las que debe responder la encuesta, normalmente están son superiores a las posibilidades del estudio y por tanto, deben reducirse estudiando si se pueden obtener de otros estudios. El paso fundamental es delimitar la población objeto de estudio y ver qué información se desea obtener, ésta esta ligada a la medida de un carácter cuantitativo o cualitativo y estos deben definirse perfectamente y de forma sencilla. Además se ha de establecer si la encuesta será o no periódica, y en caso afirmativo conservar el sistema de definiciones para posteriores encuestas.

Una encuesta por muestreo cuenta con una serie de limitaciones a tener en cuenta, entre esta limitaciones están lo medios económicos y medios humanos y materiales disponibles.

El cuestionario es el instrumento de investigación social más utilizado debido a su fácil aplicación y a la cantidad de información que puede proporcionar. Es el medio de comunicación que tenemos entre el encuestado y la persona que requiere los datos. Su elaboración debe ser un trabajo minucioso ya que será el documento de trabajo de codificadores, depuradores y personal que introduce los datos para su posterior tratamiento por los analistas. Los cuestionarios han de conjugar dos principios fundamentales:

a) Alcanzar los objetivos propuestos en el estudio:
   - ♦ Determinar los factores relevantes para el estudio
   - ♦ Transformar estos factores en la información que se ha de demandar a los entrevistados
   - ♦ Expresarlos en variables que permitan analizar la información que se obtenga

b) Resultar cómodos para los entrevistados
   - ♦ Facilitar la labor de los entrevistados, evitando preguntas en las que tengan que hacer algún trabajo de investigación o reflexión
   - ♦ Debe estar estructurado en secciones y escalonar las preguntas según nivel de dificultad (de menos a más) manteniendo en interés del entrevistado.

- El vocabulario debe ser adecuado para la persona entrevistada.

La longitud del cuestionario es una cuestión importante ya que se ha demostrado que a mayor longitud menor fiabilidad en la respuesta, por tanto, se debe evitar incluir preguntas que no sean necesarias. Los tipos de preguntas de un cuestionario se clasifican en dos: abiertas y cerradas. Las preguntas abiertas son de respuesta libre y las cerradas tienen diferentes opciones de respuesta y deben ser exhaustivas y excluyentes. Las primeras son más difíciles de codificar que las segundas. Una vez delimitada la información que se pretende recoger mediante el cuestionario, el tipo de preguntas,… suele llevarse a cabo una prueba piloto denominada pretest, con el fin de comprobar su funcionamiento. Hay ocasiones en las que es necesario modificar algunas preguntas, modificar su enunciado,…

2. Diseño de la muestra:

a) Plan de muestreo: Formación o actualización del marco (En la práctica en raras ocasiones se dispone de marcos perfectos y la correspondencia entre los elementos del marco y la población no es biunívoca), utilización de información complementaria (si anteriormente se ha realizado un estudio similar hay que ver si se utiliza el mismo método o hay que realizar modificaciones importantes), tipo de muestreo a utilizar, tamaño muestral, selección de las unidades…

b) Métodos de estimación: Tipos de estimadores y fórmulas para la estimación de errores del muestreo. Tratamiento de la falta de respuesta.

3. Trabajo de campo: Procedimientos de recogida de datos, formación de entrevistadores.

La recogida de datos debe hacerse dentro de un sistema coordinado para la gestión de los mismos. A este sistema se le denomina genéricamente Red de Campo, está formado por un conjunto de personas que facilitan la recogida de información. La estructura de la red de campo puede variar dependiendo del tipo de encuesta o del organismo que la realice, pero básicamente podemos señalar las siguientes:

- Entrevistadores: realizan la entrevista según unas instrucciones previas.
- Coordinadores: tienen a su cargo un grupo de entrevistadores
- Inspectores: se encargan de detectar fallos en la recogida de datos
- Responsables: aquellos investigadores o directores de los que depende la red de campo en su conjunto.

- Codificadores: dependiendo del sistema de recogida de la información será necesario un grupo de personas que convierta la información marcada por el entrevistador en códigos aptos para su posterior tratamiento informático.

4. Proceso de datos

En el procesamiento de datos se incluyen varias fases que en la mayoría de las ocasiones se solapan. Estos procedimientos incluyen la codificación, , la grabación, la depuración, ajuste de la no respuesta, etc. Hay una fase importante que consiste en comprobar si la muestra ha sido respetada y se han recibido todos los cuestionarios previstos en la muestra inicial.

La codificación consiste en representar las posibles respuestas por un código (un número) para facilitar el tratamiento informático. No debe ser ambigua y se deben codificar las respuestas comunes a todas las preguntas con el mismo código (Por ejemplo NS: 888, NC: 999).

Respecto a la depuración es una fase del procesamiento de datos que se mezcla con la anterior ya que muchas veces los codificadores detectan y corrigen errores antes de ser grabados.

El control de grabación es el paso de información a un soporte apto para ser tratado informáticamente. Esta fase no se lleva a cabo en sistemas de recogida CAPI y CATI. Existe un sistema de grabación inteligente si el software que se utiliza en los equipos de grabación detecta errores al introducir las respuestas.

El análisis de los resultados consiste en aplicar las técnicas estadísticas posibles y necesarias para poder extraer conclusiones fiables y significativas de los datos obtenidos. El análisis irá en función de los objetivos que se marquen.

5. Evaluación y presentación de resultados: Diferencias entre el diseño teórico y el práctico. Comparación con fuentes externas.

La mera publicación de los resultados de una encuesta, no informa sobre la complejidad de las operaciones necesarias para llevarse a cabo. La descripción de las operaciones que han sido necesarias para llevarla a cabo es la única forma de conocer la calidad de la encuesta y sus estimaciones. Con este fin es necesario presentar dos tipos de informe, el informe técnico y el informe resumido.

El informe técnico debe contener información detallada sobre fuente de datos, conceptos, clasificaciones, metodología, etc

El informe resumido va dirigido al usuario general y se debe presentar en cada difusión de la encuesta. Debe contener, como mínimo, la siguiente información:

- Fuente de datos, definiciones, clasificaciones,
- Cobertura de la encuesta, idoneidad del marco utilizado,
- Descripción de los métodos de selección y estimación,
- Tasas de respuesta y su definición,
- Error de muestreo y su interpretación.

1.6. Formas de administración: Personal, correo, telefónica, internet

Una vez seleccionado el individuo objeto de la entrevista, esta se podrá realizar sobre diversos métodos.

- Observación directa: es la única solución posible al problema de que la unidad informante no pueda dar la información que de él se requiere. Por ejemplo si se quiere medir el nivel de glóbulos rojos en sangre de un conjunto de pacientes a los que se les ha administrado un tratamiento.
- Utilización del teléfono: el teléfono puede utilizarse de varias formas, para establecer una cita entre el respondiente y el entrevistador, como recordatorio a los no-respondientes en una encuesta por muestreo…
- Utilización del correo ordinario: en este tipo de encuestas el respondiente sólo posee una serie de instrucciones que pueden ayudarle a completar el cuestionario.
- Entrevista personal: es el método más utilizado en las encuestas a gran escala. Aunque es costoso y el entrevistador puede introducir sesgos y errores es el más utilizado en investigación social.
- Encuestas on-line: En estos momentos están adquiriendo gran desarrollo las encuestas mediante correo electrónico o Internet. Presentan algunas ventajas, como su reducido coste y la posibilidad de incluir preguntas con diseños mucho más versátiles que en los otros modos de administración. Sin embargo, tienen el inconveniente de que no toda la población es usuaria de Internet, y tampoco existe listados de direcciones de correo electrónico o de usuarios.

Dentro de las entrevistas, una vez seleccionado el individuo objeto de la misma, esta se puede realizar mediante diversos métodos:

- PAPI (Paper Assisted Personal Interviewing): Entrevista clásica sobre papel

- CAPI (Computer Assisted Personal Interviewing): Los entrevistadores van provistos de PDA u ordenadores portátiles. Este método evita errores muy frecuentes ya que respeta el flujo del cuestionario (preguntas filtro), respeta los códigos de respuesta, la grabación es automática. Como desventaja hay que señalar el coste de su puesta en práctica.
- CATI (Computer Assisted Telephone Interviewing): es similar al método anterior, salvo que las entrevistas se realizan por teléfono. Actualmente presenta el inconveniente de que no todos los hogares poseen teléfono fijo y para los teléfonos móviles no existe un listado de los mismos.

2. PRINCIPALES TIPOS DE MUESTREO

2.1. Muestreo aleatorio simple

Introducción

El muestreo aleatorio simple es el tipo de muestreo en el que se basan todos los demás tipos de muestreo. Consiste en numerar las unidades de 1 a N (siendo N el tamaño de la población), y extraer una serie de n números aleatorios. Las unidades correspondientes a esos números serán las que formen parte de la muestra.

Cada unidad tendrá una probabilidad de n/N de aparecer en la muestra.

El muestreo aleatorio simple se puede realizar con o sin reemplazamiento. Si es sin reemplazamiento, una unidad que se ha escogido para pertenecer a la muestra no puede repetirse, mientras que si se realiza con reemplazamiento la unidad sí puede repetirse. Aunque las expresiones de sus estimadores y características son distintas, cuando la población es grande las diferencias son despreciables. Dado que las expresiones son más sencillas en el muestreo con reemplazamiento, utilizaremos estas, para no extendernos, aunque la mayor parte de las veces en las encuestas sociales se utilice el muestreo sin reemplazamiento.

Estimadores

Incluiremos los estimadores de medias y proporciones, dejando para su consulta en los manuales el resto de estiamdores.

El estimador de la media poblacional es la media muestral, $\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$

El estimador del error estándar de la media es $\hat{s}_y = \frac{s}{\sqrt{n}}\sqrt{1-f}$

En esta expresión s es la desviación típica muestral, y f la fracción de muestreo, es decir, n/N. Si la fracción de muestreo es muy pequeña, como suele ocurrir cuando las poblaciones son grandes, se puede prescindir de ella.

El estimador de una proporción es la proporción muestral $\hat{P} = p$

El estimador del error estándar de una proporción es: $\hat{s}_p = \sqrt{\frac{1-f}{n-1}\,p(1-p)}$

Al igual que en el caso anterior, si la fracción de muestreo es muy pequeña se puede prescindir de ella en la expresión.

Intervalos de confianza

Cuando se utilizan estimadores derivados de un muestreo es importante tener en cuenta que los resultados obtenidos no son exactos, sino que existe un margen de error. Para expresar la precisión de los estimadores se utilizan los intervalos de confianza.

El intervalo de confianza para un estimador $\hat{y}$ al nivel de confianza (1- α)% es el intervalo alrededor de $\hat{y}$ en el cual el verdadero valor del parámetro poblacional y se situará con una probabilidad 1-α.

Para calcular el intervalo de confianza hace falta tener en cuenta la distribución muestral del estimador. En el caso de la media muestral, el teorema central del límite nos permite utilizar habitualmente la aproximación normal, quedando el intervalo de confianza para $\bar{y}$ como

$$\bar{y} \pm t_\alpha \hat{s}_{\bar{y}}$$

Siendo en esta expresión t_α el valor que alcanza la distribución normal en el punto α

La expresión $t_\alpha \hat{s}_{\bar{y}}$ se denomina habitualmente nivel de error ε de la estimación, y es la que suele aparecer en las fichas técnicas de las encuestas.

Esta aproximación no es buena cuando las muestras son pequeñas y las distribuciones son muy asimétricas, por lo que en estas circunstancias debe buscarse otras técnicas de cálculo de intervalos de confianza.

El intervalo de confianza para una proporción p al nivel de confianza (1- α) se calcula de forma similar, siendo la aproximación normal $\hat{p} \pm t_\alpha \hat{s}_p$.

Cuando la proporción es muy desigual por ejemplo, mayor que 0,7 o menor que 0,3, estamos ante un caso claro de distribución muy asimétrica, por lo que en este caso solo se recomienda utilizar esta aproximación con muestras muy grandes (al menos mayores que 1.000), ya que de lo contrario el sesgo puede ser muy importante.

El nivel de confianza más utilizado en las ciencias sociales es el 95%, resultando $t_\alpha = 1{,}96$, y quedando la expresión $\hat{p} \pm 1{,}96\ \hat{s}_p$

Tamaño de la muestra

Cuando se realiza un muestreo habitualmente surge la cuestión de cual debe ser el tamaño de la muestra que se va a realizar. Para ello es necesario saber qué precisión necesitamos, y con qué probabilidad queremos que nuestro estimador esté dentro del intervalo de confianza que se obtenga con esa precisión.

Dado ε (error máximo tolerable) y 1-α (nivel de confianza) el tamaño muestral para la estimación de la media muestral en poblaciones grandes es $n = \frac{t_\alpha^2 S^2}{\varepsilon^2}$.

El problema de esta expresión es que S^2 (la varianza de la media) es normalmente desconocida, por lo que para calcularlo es necesario realizar una estimación de ella.

En el caso de una proporción la expresión sería $n = \frac{t_\alpha^2 p(1-p)}{\varepsilon^2}$.

En este caso, si no conocemos el valor de p con antelación, podemos calcular el tamaño muestral considerando el máximo que puede alcanzar la varianza de la proporción, que se alcanza en p=0,5. Entonces la expresión queda $n = \frac{t_\alpha^2 0{,}25}{\varepsilon^2}$.

Algunos de los tamaños muestrales que se deducen de esta expresión para el nivel de confianza del 95% son los siguientes.

ε	n
1%	9.604
2%	2.401
3%	1.067
4%	600
5%	384

2.2. Muestreo estratificado

Introducción

En el muestreo estratificado se realiza primero una partición de la población en subpoblaciones que se denominan estratos, y dentro de cada estrato se realiza el muestreo de forma independiente. Una condición que es requerida para los estratos es que su población debe ser conocida.

Principales utilidades del muestreo estratificado:

- Sirve cuando se quiere obtener una precisión distinta para cada subpoblación. De esta forma se puede controlar qué muestra pertenece a cada estrato, y así controlar su precisión.

- Se utiliza también cuando es necesario plantear distintas tácticas de muestreo según las subpoblaciones.
- Si los estratos que se utilizan son más homogéneos que la población, la utilización del muestreo estratificado permite ganar precisión frente al aleatorio simple.

Afijación

Afijación es la forma de realizar el reparto de la muestra en cada uno de los estratos. Algunos tipos de afijación son:

- Proporcional: La muestra es proporcional a la población. Esta afijación es la más eficiente para la población total si no tenemos ninguna condición que la contradiga, ni más información sobre costes o varianzas.
- Uniforme: La muestra es la misma en todos los estratos. Esta afijación se utiliza si tenemos un objetivo de precisión igual para todos los estratos, independiente de su población. El problema es que es menos eficiente para las estimaciones totales, por lo que con el mismo tamaño de muestra puede aumentar mucho el nivel de error total.
- De compromiso: En esta afijación se utilizan unos tamaños maestrales intermedios entre los que resultarían de aplicar las dos anteriores. Se utiliza mucho si se quiere controlar la precisión tanto para la población total como para cada estrato.
- Óptima: Tiene en cuenta para calcular los tamaños de los estratos las varianzas de los estratos, y los costes de realización de la muestra en cada uno de ellos. Es útil tenerla en cuenta si las varianzas son muy distintas de unos estratos a otros, y si los costes de realizar el estudio son muy distintos de unos estratos a otros.

Estimadores

El estimador para la media en el muestreo estratificado es la media de las medias de cada estrato ponderada por la población de los estratos: $\hat{\bar{Y}}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h$

En el caso de la afijación proporcional, el estimador coincide con la media muestral.

El estimador del error estándar de la media es: $\hat{s}_{st} = \sqrt{\sum \frac{s_h^2}{n_h} W_h^2 (1 - f_h)}$.

En esa expresión f_h es la fracción de muestreo del estrato *h*, que en estratos con población grande se puede omitir.

Tamaño muestral

Para el cálculo del tamaño muestral en un muestreo estratificado es necesario tener en cuenta el tipo de afijación y los objetivos de error.

Si la afijación es proporcional prácticamente se puede usar el mismo procedimiento que en el muestreo aleatorio simple. Si existe objetivos de error para los estratos, por los que se va a utilizar una afijación uniforme, la muestra se tendrá que calcular para cada uno de los estratos de forma independiente. Finalmente, si la afijación es de compromiso es preferible calcular primero el tamaño para el objetivo de cada estrato y después aumentar la muestra en los estratos necesarios para cumplir el objetivo en la población total.

2.3. Muestreo sistemático

Este muestreo consiste en ordenar a la población de acuerdo con una variable. Tomamos una muestra aleatoria entre las primeras $k = \frac{N}{n}$ unidades, y tomamos las n siguientes a intervalos de amplitud k.

El muestreo sistemático es más preciso que el m.a.s. si la variable por la que se ordena está relacionada con la que nos interesa, pero se pueden presentar sesgos si la variable de ordenación tiene comportamientos secuenciales.

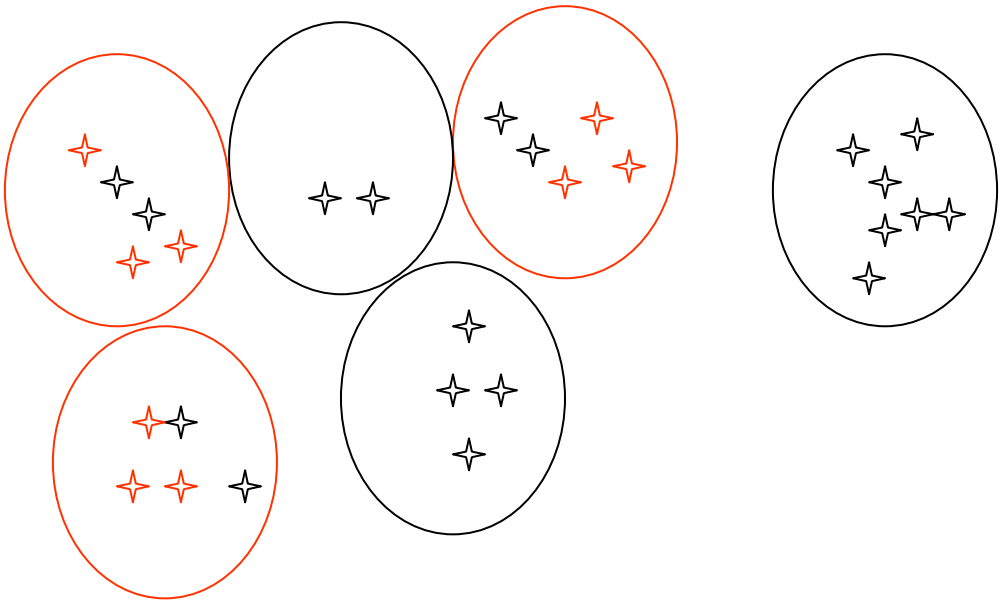
El muestreo sistemático se puede considerar un caso particular de muestreo estratificado, donde cada tramo de amplitud k estrato.

2.4. Muestreo por conglomerados

Introducción

Este tipo de muestreo consiste en usar unas unidades intermedias, llamadas conglomerados, y muestrearlos. Dentro de cada conglomerado existe una parte de las unidades finales. Se puede incluir en la muestra a todas las unidades de los conglomeradas elegidas, o solo a una muestra de ellas, tratándose en el segundo de los casos de un muestreo por conglomerados con submuestreo, o muestreo en dos fases.

En la figura siguiente tenemos un ejemplo de muestreo por conglomerados, estando en rojo los conglomerados y unidades finales seleccionadas.



El muestreo por conglomerados es muy utilizado en las encuestas sociales, ya que es la solución más adecuada si no se cuenta con un listado de unidades, o es muy costoso conseguirlo, pero sí del listado de agrupaciones, o bien si aunque se disponga de dicho listado es muy difícil o costoso acceder a las unidades elegidas de forma aleatoria, pero sí es más fácil el acceso a ellas agrupadas por los conglomerados (por ejemplo, por motivos de distancia).

El muestreo por conglomerados en general aumenta el error, ya que los conglomerados suelen ser más homogéneos que la población. El error aumenta menos si se toma un mayor número de conglomerados, disminuyendo por tanto el número de unidades muestreadas en cada conglomerado.

Probabilidad de elección de la unidad última y número de unidades

En el muestreo por conglomerados es conveniente, salvo una razón poderosa en contra, que la probabilidad de elección de la unidad última sea la misma siempre, ya que simplifica la forma de los estimadores. Hay dos formas de conseguir esto:

a) Elegir a los conglomerados con probabilidad proporcional al tamaño, y muestrear después el mismo nº de unidades en todos.

b) Elegir a los conglomerados con igual probabilidad, y muestrear un número en cada conglomerado un número de unidades proporcional al tamaño

Estimadores y errores

Las expresiones en muestreo por conglomerados con submuestreo son complejas, por lo que se escapa a los objetivos de este documento introductoria. Sin embargo, si las probabilidades de las unidades últimas son las mismas, los estimadores son los mismos del m.a.s.

Para el cálculo de errores se hace necesario utilizar métodos complejos, por lo que es preferible acudir a programas adecuados para realizarlo (SPSS Muestras Complejas, WesVar, STATA, SAS,…).

3. Ejemplo: muestreo para encuesta social a la población andaluza

Como ejemplo, vamos a mostrar el proceso de decisiones que se ha tomado en el IESA para la realización de una muestra concreta, dirigida a la población residente en Andalucía mayor de 18 años.

3.1. Forma de administración y marcos muestrales

Formas de administración

En primer lugar tenemos que decidir la forma de administración. En principio se considera solo la administración presencial o telefónica. Consideramos la duración del cuestionario (la telefónica debe usarse solo en cuestionarios cortos), el coste, y el tiempo que se tiene para realizar el trabajo de campo (en estos dos aspectos es mejor la telefónica).

También hay que considerar la importancia que le vamos a dar a los posibles sesgos introducidos por el marco muestral de las encuestas telefónicas.

Marcos muestrales en encuestas telefónicas

En las encuestas telefónicas los marcos maestrales existentes son solo de teléfonos fijos (las bases de datos procedentes de los listines telefónicos). En estos momentos, se constata que hay una parte importante de la población, sobre todo jóvenes, donde se ha sustituido el teléfono fijo por el móvil, por lo que el sesgo producido en este tipo de encuestas puede ser importante.

Se están realizando ya encuestas a teléfonos móviles para solucionar este problema, pero tiene el inconveniente de que no existe un listado de números, y tampoco hay forma de saber a qué lugar de España pertenece cada número, lo que complica el muestreo si no es a la población nacional, en caso de realizarse con elección aleatoria de números de teléfono.

Marcos muestrales en encuestas presenciales

Para la población general el único marco muestral existente es el Padrón de Habitantes, que realiza el INE en colaboración con los ayuntamientos. Por desgracia, es difícil poder utilizar este marco para realizar encuestas, ya que las condiciones de acceso son bastante estrictas por motivos de confidencialidad.

Decisión

En nuestro caso, dado que el cuestionario es extenso, elegimos la encuesta vía presencial. Como no disponemos del Padrón de Habitantes, debemos recurrir a una estrategia basada en la utilización de las secciones electorales como unidad intermedia de muestreo.

3.2. Secciones electorales

El Instituto Nacional de Estadística, en colaboración con los ayuntamientos, ha dividido todo el territorio nacional en secciones electorales de un tamaño más o menos homogéneo, entre 1.500 y 2.000 habitantes, salvo municipios más pequeños en los que la sección coincide con el municipio.

De estas secciones se dispone de su población actualizada, callejero y forma geográfica, por lo que son una buena unidad para acotar la distribución de la población, y muy usadas en este tipo de encuestas como conglomerados, incluso si se dispone de registro de población, para ahorrar excesivos costes de desplazamiento.

3.3. Rutas aleatorias

Si no se dispone del listado de población una posibilidad es la utilización de rutas aleatorias, que pueden considerarse un muestreo sistemático de los hogares de la sección. Cada organización tiene sus propias normas para realizar una ruta aleatoria. En el caso del IESA, son las siguientes:

El *punto de partida* para el entrevistador será la calle de la sección electoral que se le señale en las instrucciones de cada estudio. De la calle seleccionada, se partirá del número de portal/vivienda más bajo que aparezca y, dentro de éste, del 1er. Piso y 1ª letra.

El *recorrido* de la sección se hará con las siguientes normas:

A. Para *edificios de más de una vivienda*, un portal cada tres (uno sí, dos no). En edificios con dos o más escaleras se tomarán como portales cada una de ellas. Dentro del portal seleccionado se podrá realizar una entrevista cada 12 hogares o fracción.

B. Para viviendas unifamiliares, un portal de cada 5 (uno sí, cuatro no). En este caso, el segmento para la selección de la vivienda es de 5 hogares.

C. No se respetarán los segmentos en los casos de viviendas diseminadas

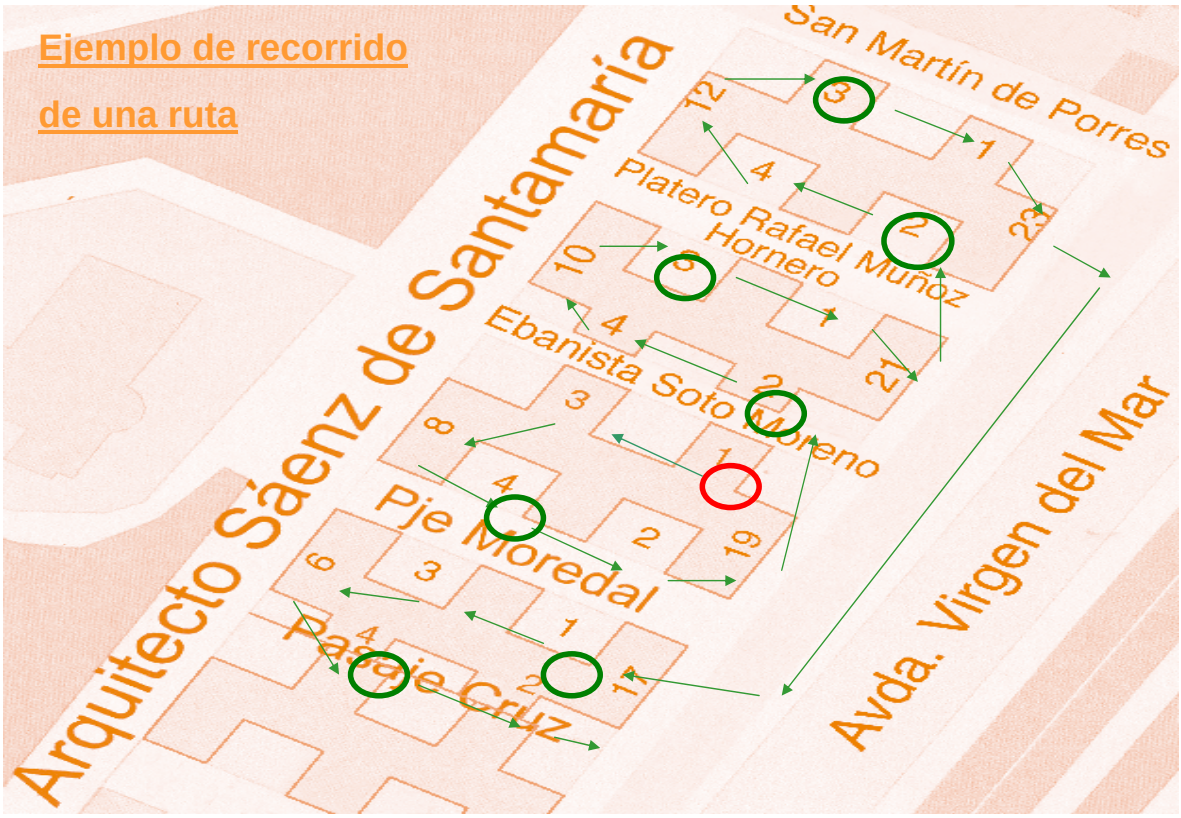
Ejemplo de recorrido de una ruta

La *ruta* es el itinerario que debe seguir el encuestador/a para realizar un número determinado de entrevistas y tiene como límites una sección electoral.

Para poder realizar la ruta, se proporciona a los encuestadores/as el listado de calles que componen esa sección electoral.

LISTADO DE CALLES DE LA SECCIÓN ELECTORAL CORRESPONDIENTE AL DISTRITO 5 SECCIÓN 12 DE CÓRDOBA

Tipo de vía	*Nombre de la vía*	*Impares*	*Pares*
CALLE	Arquitecto Sáenz Santamaría	del 1 al 1	Del 6 al 12
PSAJE	De la Cruz	Del 2 al 4	
PSAJE	Ebanista Soto Moreno	Todos	Todos
PSAJE	Del Moredal	Todos	Todos
CALLE	Platero Rafael Muñoz Hornero	Del 1 al 3	Del 2 al 4
CALLE	San Martín de Porrés	Del 1 al 3	
AVDA	De la Virgen del Mar	Del 17 al 23	



3.4. Reintentos y cuotas

Como se ha dicho, en estas condiciones la ruta se puede considerar un muestreo sistemático, por lo que la elección de las personas es perfectamente probabilística. El problema aparece cuando en el hogar donde se pretende entrar no está la persona que corresponde entrevistar. En ese caso, se debería volver al hogar una y otra vez hasta que se encuentre. Problemas prácticos, de tiempos, costes y otras consideraciones hacen que en la práctica no se suela hacer esto, y si en una vivienda no hay nadie se pasa a la siguiente de la ruta. Aquí entramos por tanto en un muestreo no probabilístico, ya que no sabemos la probabilidad que tiene una persona de estar en su hogar cuando se realiza la ruta.

Para controlar los sesgos que puede llevar implícita la utilización de este tipo de muestreo en esta fase última se suelen utilizar cuotas en las que se señala el número de personas que se pueden entrevistar en cada sección de acuerdo a una serie de variables que pensamos que pueden estar relacionadas con nuestro estudio. Las más utilizadas son la edad y el sexo.

Instrucciones para la selección de la persona a entrevistar

El total de individuos a entrevistar que compone la muestra se fragmenta en una serie de hojas de cuotas de sexo y edad, una para cada punto de muestreo, de tal forma que la unión de todas ellas coincida con la distribución muestral.

Ejemplo de cuotas de edad y sexo

Nº DE RUTA : 31
PROVINCIA : 14 – CÓRDOBA
MUNICIPIO : 21 - CÓRDOBA
DISTRITO : 5 SECCION : 12
TOTAL DE ENTREVISTAS : 8

EDAD	VARÓN	MUJER
18 A 29	1	2
30 A 44	1	1
45 A 59	1	
60 O MÁS	1	1

3.5. Estratificación y determinación del tamaño de la muestra

En este tipo de muestreo donde se utilizan secciones es muy habitual la utilización de estratificaciones de las secciones.

A menudo se suele utilizar las provincias como estrato, sobre todo si se requiere una precisión determinada para ellas.

Otras variables de estratificación pueden ser el tamaño del municipio, o variables sociodemográficas derivadas del Censo de Población.

En nuestro caso los requisitos que nos planteaban era la obtención de un error de estimación máximo para las proporciones del 5% para cada provincia y 1,7% para el total de Andalucía. Por tanto, se decide realizar una afijación de compromiso, donde el mínimo de entrevistas por

estrato va a ser de 384, aumentando en las provincias mayores el número de entrevistas para obtener la precisión requerida para el total[1].

De esta manera, el reparto de entrevistas por estrato queda

Provincia	Muestra
Almería	384
Cádiz	430
Córdoba	384
Granada	384
Huelva	384
Jaén	384
Málaga	540
Sevilla	650
Total	3540

Además de las provincias se utilizarán también estratos de tamaño de municipio, como variable que está relacionada con nuestro objeto de estudio. Utilizando cuatro tramos de tamaño municipal resulta un total de 32 estratos.

3.6. Elección aleatoria de los conglomerados

Como se ha dicho en cada estrato se elegirán las secciones que formarán parte de la muestra.

Hay que decidir el número de unidades maestrales finales (personas) que se van a incluir en cada sección. Como se ha dicho, es conveniente seleccionar cuantos más conglomerados mejor, para aumentar la dispersión y así diminuir el error muestral. La limitación de coste más importante es que para sección será necesario un desplazamiento. Por tanto se decide ajustar el número de unidades en cada sección al número de entrevistas que se supone que será posible que un entrevistador realice en una jornada de trabajo. En nuestro caso se fijará en 7 entrevistas. Por tanto, tendremos que seleccionar 3540 entrevistas entre 7 igual a 505,7. Redondeando por exceso, quedan 506 secciones.

3.7. Cuotas

Como se ha dicho, es conveniente utilizar cuotas para asegurar un cierto control en la eleccion de las personas entrevistadas. Las cuotas que se utilizarán serán el grupo de edad y el sexo,

[1] En realidad hay que tener en cuenta el efecto del diseño derivado del uso de conglomerados, que escapa a los contenidos de este curso

por considerarse que son las más fáciles de responder por los entrevistados, que guardan relación con nuestros objetivos, y además se dispone de información sobre ellas en la población.

Queda la labor de repartir las cuotas entre las secciones elegidas. Esto se realiza siguiendo los siguientes criterios:

a) El reparto en cada sección debe ser lo más parecido posible a la proporción que existe realmente en la población.

b) En cada estrato el reparto por cuotas de la muestra debe ser igual al de la población.

Bibliografía:

Rueda García, M.M. y Arcos Cebrián, A., (1998) *Problemas de Muestreo en Poblaciones Finitas,* Grupo Editorial Universitario.

Azorín, F. y Sánchez-Crespo J.L., (1994) *Métodos y Aplicaciones del muestreo,* Alianza Universidad Textos, Madrid

Fernández García, F.R. y Mayor Gallego, J.A., (1994) *Muestreo en Poblaciones Finitas: Curso Básico,* PPU, Barcelona

Cochran, W. G., (1963) *Sampling Techniques,* John Wiley & Sons, EEUU