Tema 1

Introducción a la Teoría de Muestras

Contenido

1.1. Introducción

El propósito de la Estadística es extraer conclusiones acerca de la naturaleza de una población. Cada vez es mayor la proporción de investigadores, en las más diversas disciplinas científicas, que realizan análisis estadísticos de datos como procedimiento formal para llegar a conclusiones o apoyar a procesos de decisión sobre las hipótesis de la investigación.

Para alcanzar su objetivo la metodología estadística puede desglosarse en varias etapas:

Planteamiento del problema. En esta etapa deben indicarse el tipo de inferencias que quieren hacerse sobre la población y las características que interesan de la misma.

Planificación del experimento. En ella se fijan las pautas para la realización del experimento, de acuerdo con las inferencias que se quieren hacer.

Selección de la muestra. Se determina la secuencia de elementos de la población sobre los que va a realizarse el experimento, para obtener la muestra base de las inferencias.

Obtención de los datos. En esta etapa se obtienen los datos correspondientes a las mediciones u observaciones en la muestra.

Tabulación de los datos. Se recogen los datos de forma que puedan utilizarse para las inferencias pertinentes.

Análisis de los datos. Se trata de extraer y condensar la información contenida en los datos de la muestra que pueda ser relevante para las inferencias.

Interpretación de los resultados. En esta etapa se hacen inferencias sobre la población a partir de los datos disponibles.

Todo estadístico sabe que la adquisición de la información que necesita es un trabajo ingrato, difícil y costoso. La recogida de datos es un punto importante en todo estudio estadístico, porque el método de análisis mejor elaborado tiene poco provecho si se aplica a datos falsos, tomados en malas condiciones.

Las Técnicas de Muestreo se ocupan de determinar, para cada estudio concreto, cuál es el procedimiento óptimo para recopilar la información que se precisa.

1.2. ¿Por qué el muestreo?

Existen dos estrategias posibles para la recopilación de datos:

- examinar todas las unidades de la población, es decir, realizar un censo, y
- examinar, según unos planes establecidos con anterioridad, ciertas unidades de la población (muestra), y suponer que los resultados obtenidos son representativos de toda la población.

A pesar de la importancia del muestreo en la investigación, existen dos supuestos generalizados que mueven a investigadores de diversas ciencias a descuidar la fase del muestreo en sus estudios:

1. **Supuesto de uniformidad.** Dado que las características poblacionales son más o menos constantes en toda su extensión, cualquier porción es representativa del total. Esta es la justificación para las muestras de sangre en los análisis de laboratorio: la composición del líquido es más o menos constante en todo el cuerpo.
2. **Supuesto de disposición aleatoria.** La población no es uniforme en cuanto a los valores de la/s variable/s que interesa/n en el estudio, pero éstos se reparten aleatoriamente en toda la extensión poblacional. Por esta razón, cualquier porción de la población será una muestra aleatoria tan válida como si se hubiera cuidado muy especialmente el componente probabilístico en la selección.

Parece, no obstante, que estos argumentos dan buenos resultados en ciencias como astronomía, física y química, si bien es poco aconsejable asumirlos en las ciencias sociales, medicina y biología. Es más, en general mantener el supuesto de homogeneidad o uniformidad, es un proceder lamentable en muestras que incluyen personas.

La conveniencia del muestreo frente a censos o investigaciones exhaustivas está totalmente justificada en las siguientes situaciones:

- Cuando la población sea tan grande que el censo exceda de las posibilidades del investigador.

- Cuando la población sea suficientemente uniforme para que cualquier muestra dé una buena representación de la misma.

- Cuando el proceso de medida o investigación de las características de cada elemento sea destructivo o disminuya su valor, como ocurre al consumir un artículo para juzgar su calidad, o al determinar una dosis letal o un punto de ruptura.

Además de los casos extremos anteriores, existen otras razones que pueden hacer ventajoso el estudiar una población a partir de sus muestras:

- **Coste reducido**. Si los datos que buscamos los podemos obtener a partir de una pequeña parte del total de la población, los gastos de recogida y tratamiento de los datos serán menores. Por ejemplo, cuando se realizan encuestas previas a un referéndum, es más barato preguntar a 4.000 personas su intención de voto, que a 30.000.000. No sólo hay que considerar el coste absoluto, sino también el relativo, esto es, el coste en relación a la cantidad de información obtenida. Puede ocurrir que el aumento de información que se obtenga con un censo no compense su mayor coste.

- **Mayor rapidez**. Disponiendo de ciertos recursos puede obtenerse mediante muestras información más rápida, frecuente y detallada, lo que aumentará su utilidad, sobre todo para fenómenos dinámicos, evolutivos. Estamos acostumbrados a ver cómo con los resultados del escrutinio de las primeras mesas electorales, se obtiene una aproximación bastante buena del resultado final de unas elecciones, muchas horas antes de que el recuento final de votos haya finalizado;

- **Calidad**. El muestreo exige, en comparación con la realización de censos y estudios exhaustivos, menos trabajo material pero más refinamiento y preparación. Requiere no solamente una base adecuada en los diseñadores, sino también una cierta preparación de los entrevistadores, inspectores y supervisores.

No hay que ver una oposición entre censo y muestra. Las posibilidades ofrecidas por cada uno de ellos son complementarias. Ello explica que estos dos tipos de encuestas sean realizadas con frecuencia de forma simultánea.

1.3. Aplicaciones del muestreo

La palabra "encuesta" es familiar a la mayoría de las personas debido a la publicidad de las encuestas de opinión política. Sin embargo, las encuestas de opinión sólo representan una parte pequeña de los muestreos.

Desde tiempos muy antiguos, los dirigentes de los países o de grandes colectivos, se han interesado por la información demográfica. Desde el punto de vista militar, esta preocupación se extiende hoy al conocimiento de bienes y servicios, con el fin de prever y regular los factores de producción.

El estadístico es solicitado, junto con otras personas, para evaluar el tamaño de las poblaciones, los tipos de bienes de los que disponen, y sus comportamientos frente a los problemas económicos y sociales. Un cuidado especial se le aplica al estudio de los problemas ligados al trabajo: la importancia de la población en paro, datos sobre las horas de trabajo, niveles de salarios, efectivos de las diversas ramas de actividad, etc.

Se encuentran también bastantes aplicaciones en los estudios de agricultura, tanto en los países desarrollados, como en los en vías de desarrollo. El tipo de datos a recoger es fácil de imaginar: superficies cultivadas, tipos de cultivos y superficies dedicadas a ellos, modos de explotación, rendimientos, ...

La actividad conocida como investigación de mercados depende en gran parte del uso del muestreo. Las estimaciones de la magnitud del auditorio de diferentes programas de radio y de televisión y de los lectores de periódicos y revistas se obtienen continuamente. Los fabricantes y detallistas quieren conocer las reacciones de la gente hacia un nuevo producto o nuevos métodos de presentación: sus quejas en relación a los productos antiguos y sus razones para preferir un producto frente a otro.

El comercio y la industria utilizan el muestreo en un intento de aumentar la eficacia de sus operaciones internas. Las actividades de control de calidad y muestreo de aceptación se fundamentan en decisiones que presuponen que los datos de la muestra son válidos para la producción completa. El muestreo de los registros de transacciones comerciales (cuentas, nóminas, personal, ...) usualmente más sencillo que el de personas, pueden proporcionar información útil, rápida y económica. También se pueden obtener ahorros, a través del muestreo, en la estimación de inventarios, en estudios de la condición y tiempos de vida de maquinaria y equipo, en la inspección de la exactitud y rapidez de trabajo del empleado, etc.

En el campo de la contabilidad y la auditoría ha surgido un nuevo interés en la adaptación de nuevas técnicas a problemas particulares en este campo.

Los campos nombrados conciernen principalmente a las estadísticas nacionales y de empresas. Pero los muestreos se utilizan en casi todos los campos científicos: ciencias de la educación, la salud, biología, meteorología, ecología, y por supuesto, en los estudios de la opinión pública y los escrutinios electorales, que han hecho tanto para generalizar la técnica de muestreo frente a los ojos del público.

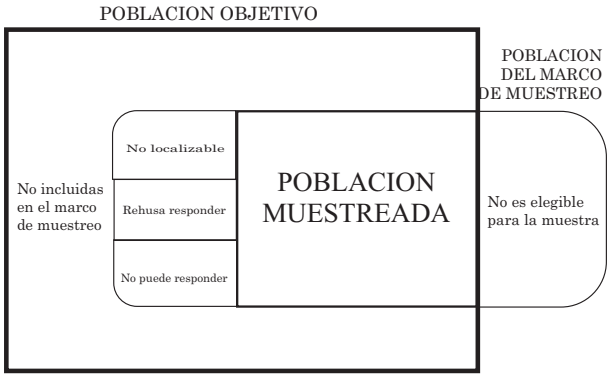
1.4. Conceptos básicos

Denominamos POBLACIÓN a la colección de elementos sobre la que quiere examinarse el comportamiento del experimento considerado. Los elementos de la población se denominan INDIVIDUOS. Llamamos MUESTRA al conjunto de individuos de la población de los cuales se obtiene información, si no se puede obtener de todos los individuos que la componen.

La población que se intenta investigar o POBLACIÓN OBJETIVO y el conjunto que realmente investigamos y que llamaremos POBLACIÓN MUESTREADA, muchas veces no coinciden ya que existen omisiones, duplicaciones y unidades extrañas. Por otro lado, la información no podrá obtenerse de algunas unidades por diferentes motivos, como la inaccesibilidad para unos medios dados, la negativa a colaborar o las ausencias.

En la etapa de selección de una muestra a partir de una población los individuos pueden extraerse de uno en uno o agrupados, dependiendo de la técnica de muestreo que se haya seguido. Así, por ejemplo, en una población humana la selección puede hacerse persona a persona, familia a familia, edificio a edificio, etc. Una UNIDAD MUESTRAL será un individuo o conjunto de individuos que se seleccionan en una única extracción. Las unidades muestrales distintas no deben tener elementos comunes y deben cubrir toda la población. Recibe el nombre de MARCO DEL MUESTREO el conjunto de todas las unidades muestrales consideradas.

Por ejemplo, una encuesta telefónica sobre intención de voto. No todas las familias tiene teléfono, de modo que varias personas de la población objetivo de posibles votantes no tendrán un número telefónico en el marco de muestreo. En algunas casas con teléfono, los residentes no están registrados para votar y, por lo tanto, no son elegibles para la encuesta. Algunas personas elegibles en la población del marco de muestreo no responden porque no pueden ser localizadas, algunas se niegan a contestar a la encuesta y algunas podrían estar enfermas o incapacitadas para responder.



Las poblaciones que vamos a considerar deben ser tales que sobre sus individuos puedan observarse ciertos rasgos o magnitudes, que varían de unos individuos a otros y que llamaremos VARIABLES. La información usual que se quiere recabar sobre la población se refiere a algunas características de estas variables, a las que denominaremos CARACTERÍSTICAS POBLACIONALES. Entre éstas, las más habituales en los estudios de Técnicas de Muestreo son: la media poblacional, el total poblacional y la proporción poblacional.

Cuando se extrae una muestra, los datos obtenidos a partir de ella nos permiten inferir unos valores aproximados de la población en su totalidad. A estos valores aproximados se les denomina ESTIMACIONES, las cuales vendrán afectadas por un error que denominamos ERROR DEBIDO AL MUESTREO. Cuanto menor sea éste, diremos que mayor es la PRECISIÓN de las estimaciones.

1.4.1. Tipos de muestreo

Los procedimientos de muestreo que permiten calcular de antemano la probabilidad de seleccionar cada muestra se denominan MUESTREOS PROBABILÍSTICOS y tienen las siguientes propiedades: Si se especifica el tamaño n de la muestra que va a extraerse y se considera el conjunto de las distintas muestras posibles de tamaño n que pueden extraerse de la población por el método de muestreo considerado, un muestreo probabilístico es aquel que asigna una probabilidad conocida a cada muestra distinta del conjunto, de manera que cada una se selecciona de acuerdo con dicho esquema de probabilidades sobre el conjunto. Dentro de los métodos de muestreo probabilísticos encontramos los siguientes tipos:

- **Muestreo aleatorio**. De la población se extrae un muestra de tamaño n, dando a cada unidad la misma probabilidad de ser extraída. La muestra se puede extraer:
  - ***sin reposición (m.a.s.)***: una vez extraída una unidad, no se la vuelve a tomar en cuenta para las siguientes extracciones.

  - ***con reposición***: una unidad seleccionada en una extracción se repone en la urnaz participa en las siguientes extracciones; se pude seleccionar dicha unidad dos veces o más.

- **Muestreo sistemático**. De una población con N unidades numeradas en algún orden, para seleccionar una muestra de n unidades (siendo N = nk) se toma al azar una unidad entre las k primeras unidades, y de ahí en adelante se toma cada k-ésima unidad.

- **Muestreo aleatorio estratificado**. En una población dividida en L estratos (subconjuntos homogéneos respecto a alguna(s) característica(s) determinada(s) a priori), para seleccionar una muestra de tamaño n, se extrae una muestra en cada uno de los estratos de forma que la muestra final está compuesta por el conjunto de las submuestras ($n = n_1 + \cdots + n_L$) . El proceso de muestreo se realiza de modo independiente en cada estrato.

  La afijación (asignación) del tamaño muestral dentro de cada estrato puede ser

  - *Uniforme*. Se asigna a todos los estratos el mismo número de unidades muestrales.
  - *Proporcional*. Se asigna a cada estrato un peso proporcional al que tiene en la población.
  - *De mínima varianza o afijación de Neyman*. Consiste en determinar los $\{n_h\}_{h=1,\ldots,L}$ de forma que para un tamaño de muestra fijo, n, la varianza del estimador sea mínima.
  - *Óptima*. Consiste en determinar los $\{n_h\}_{h=1,\ldots,L}$ de forma que para un coste fijo, C, la varianza de los estimadores es mínima.

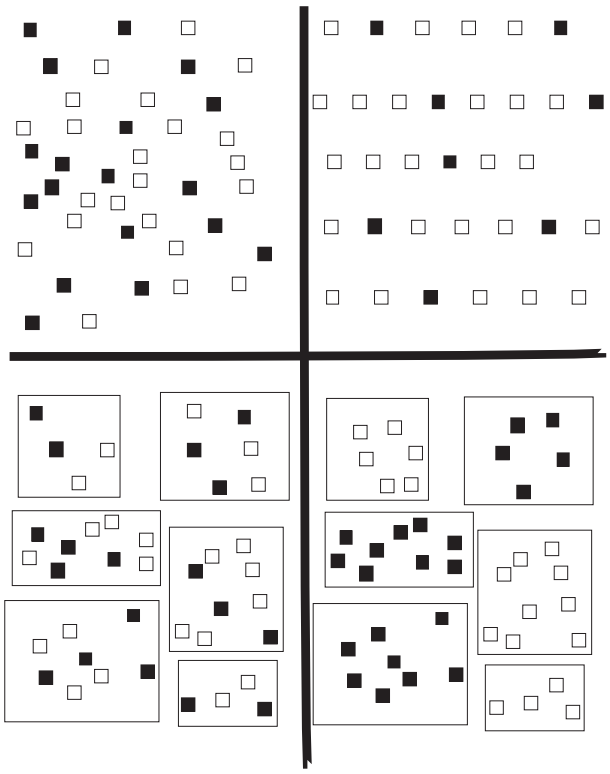
- **Muestreo por conglomerados**. En una población compuesta de M conglomerados (subconjuntos heterógeneos), para seleccionar una muestra de tamaño n, una vez seleccionados algunos conglomerados, la muestra está formada por todas las unidades que componen el conglomerado (muestreo unietápico) . Si a su vez se extraen muestras dentro de cada conglomerado seleccionado, se pasaría al muestreo multietápico.

A veces, para estudios exploratorios, el muestreo probabilístico resulta excesivamente costoso y se acude a métodos no probabilísticos. En general se seleccionan a los sujetos siguiendo determinados criterios que proporcionen muestras que sean representativas. Pueden mencionarse varias formas de MUESTREO NO PROBABILÍSTICO:

Muestreo sin norma Se toma la muestra de cualquier manera, por razones de comodidad o circunstancias o capricho. La representatividad de tal muestra puede ser satisfactoria si la población es homogénea.

Muestreo semiprobabilístico Es un procedimiento tal que el carácter probabilístico se mantiene sólo hasta un punto del proceso de selección.

Muestreo intencional u opinático La selección se lleva a cabo según criterio de autoridad. La representatividad depende de la intención u opinión de la persona que la obtiene y en este caso la composición de la muestra puede estar influenciada por sus preferencias o tendencias, incluso inconscientemente.



1.4.2. Tipos de errores

Obviamente para cualquier método de muestreo probabilístico puede hallarse la distribución de frecuencias de las estimaciones que ofrecería si se aplicara repetidamente sobre la población. También va a poder discutirse la representatividad de la muestra y la precisión de las estimaciones en términos estadísticos. Para ello se determina la DISTRIBUCIÓN EN EL MUESTREO del procedimiento de la estimación a partir de la distribución sobre el conjunto de muestras distintas.

Un criterio posible para juzgar la representatividad de un estimador es el basado en el SESGO o diferencia entre el valor esperado del estimador a los largo de las muestras distintas posibles y el valor verdadero de la característica poblacional.

En general si $\hat{\theta}$ es un estimador del parámetro θ, el sesgo se define como $b(\hat{\theta}) = E(\hat{\theta}) - \theta$. Una propiedad deseable para $\hat{\theta}$ es que sea insesgado del parámetro desconocido, es decir $E(\hat{\theta}) = \theta$.

Un criterio usual para valorar la precisión del procedimiento es determinar su ERROR CUADRÁTICO MEDIO:

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2,$$

que coincide con la varianza en el caso de ser insesgado. En general,

$$ECM(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + b(\hat{\theta})^2$$

Se llama ERROR DE MUESTREO a la raíz cuadrada de la varianza del estimador o de su error cuadrático medio, según que el estimador sea insesgado o tenga sesgo, respectivamente.

Se llama ERROR DE MUESTREO RELATIVO (o COEFICIENTE DE VARIACIÓN) de un estimador al cociente entre su error de muestro y su esperanza:

$$CV(\hat{\theta}) = \frac{\sqrt{ECM(\hat{\theta})}}{E(\hat{\theta})}.$$

Hay ocasiones en que es preferible dar en vez de una estimación puntual de un parámetro, un intervalo al que pertenezca el parámetro con una probabilidad dada. El intervalo que se calcula a partir de la muestra seleccionada se denomina intervalo de confianza del parámetro. Consideraremos dos formas diferentes de construir intervalos de confianza:

- Partiendo de la desigualdad de Tchebychev, se obtiene que si $\hat{\theta}$ es un estimador de θ
$$\left(\hat{\theta} - k\sqrt{ECM(\hat{\theta})}, \hat{\theta} + k\sqrt{ECM(\hat{\theta})}\right)$$
es un intervalo de confianza para el parámetro θ con un nivel de confianza de $\left(1 - \frac{1}{k^2}\right)$.

  Para valores númericos grandes del error del estimador, puede ocurrir que el intervalo sea prácticamente inútil por obligarnos a elegir intervalos exageradamente amplios para un coeficiente aceptable.

- Si el tamaño de muestra es suficientemente grande para que la distribución en el muestreo del estimador tienda a la normal, lo cual sólo exige que la población de origen tenga varianza finita, en virtud del Teorema Central del Límite, se puede obtener el siguiente intervalo de confianza:
$$\left(\hat{\theta} - z_{1-\alpha/2}\sqrt{ECM(\hat{\theta})}, \hat{\theta} + z_{1-\alpha/2}\sqrt{ECM(\hat{\theta})}\right)$$
donde $z_{1-\alpha/2}$ es el valor de la distribución normal $\mathcal{N}(0,1)$ que acumula una probabilidad $(1-\alpha/2)$.

Además de los errores de muestreo, existen otros ERRORES AJENOS AL MUESTREO que se presentan tanto en las encuestas como en los censos. Principalmente, se pueden distinguir las siguientes fuentes de error no muestral:

- *Error de cobertura.* Aparece cuando hay discordia entre la población muestreada y la población objetivo ya sea por defecto -omisiones- o por exceso -duplicidades y unidades extrañas-.

- *Error de selección.* Surge cuando no se respeta la selección muestral original determinada por el procedimiento de muestreo que se está considerando.

- *Error de respuesta.* Se produce cuando no es posible obtener las mediciones de interés sobre algún elemento en la muestra. Puede deberse a: la ausencia temporal del encuestado durante las horas de entrevista, negativa absoluta a colaborar, falta de conocimientos o capacidad por parte del informante, método de recogida de datos, condiciones personales y grado de adiestramiento de los entrevistadores, motivación de los encuestados, etc. Existen algunos métodos de tratamiento de la falta de respuesta.

- *Error de medición.* Se da si la información obtenida a partir de la muestra es incompleta o errónea.
- *Error de tratamiento de los datos.* Aparece cuando la edición, codificación o tabulación de los datos es defectuosa.

1.5. Objetivos de las Técnicas de Muestreo

Supongamos que tenemos una población finita cualquiera de la que se desea conocer una característica de sus elementos. En el momento en el que se decide medir sólo una parte de las unidades de la población de interés y estimar el dato desconocido a partir de dicha información, en lugar de proceder a obtener el dato en cuestión para todas y cada una de las unidades, diremos que se ha planteado un problema de muestreo.

Para dejar solucionado el problema hay que llevar a cabo una serie de tareas que se condicionan mutuamente y que hay que abordar conjuntamente:

1. Delimitar el número de unidades o elementos que han de seleccionarse.
2. Establecer la forma en que hay que efectuar la selección.
3. Determinar el modo en que se procesarán los datos para realizar la estimación.
4. Dar el procedimiento de cálculo del error que se comete en el proceso de estimación.

Cabe distinguir entre estas tareas dos tipos de cuestiones:

- Las cuestiones prácticas relativas a la elección de las unidades muestrales, realización de encuestas, elaboración de cuestionarios, preparación de los encuestadores, etc., que básicamente no son de naturaleza estadística.
- Las cuestiones propiamente estadísticas, como son la elección de procedimientos de muestreo adecuados, la elección de métodos de estimación de las características poblacionales y la legitimación de las interpretaciones de los resultados, requieren una formalización que es el objeto de la Técnicas de Muestreo.

Así, el objetivo de las Técnicas de Muestreo es estudiar procedimientos de selección de la muestra y de estimación que, con el coste mínimo posible, proporcionen estimaciones con la mayor eficacia posible.

La eficacia se traduce en la precisión de las estimaciones y en el coste del muestreo. Obviamente, poca información suele dar lugar a estimaciones poco precisas y excesiva información suele conllevar un coste muy elevado. La cantidad de información contenida en la muestra depende de su tamaño y de la variación en los datos muestrales. A su vez, esa variación depende del método de selección de la muestra.

En Técnicas de Muestreo nos vamos a ocupar de estudiar los procedimientos de selección de la muestra más adecuados según distintos criterios, los procedimientos de estimación de las características que se consideren, los procedimientos de cuantificación de la precisión de las estimaciones y los procedimientos para determinar tamaños muestrales adecuados.

Tema 2

Muestreo aleatorio simple y con reposición

Contenido

2.1. Introducción

Por **muestreo aleatorio** se entiende que en cada extracción todas las unidades disponibles en esa extracción tienen la misma probabilidad de ser elegidas.

La versión más básica del muestreo probabilístico es el **muestreo aleatorio simple** (m.a.s.), también llamado **muestreo aleatorio sin reposición** que conduce a observaciones que no son independientes y por lo tanto no se corresponde con lo que llamamos m.a.s. en Estadística Matemática.

El muestreo aleatorio con reposición da lugar a un estudio más sencillo aún y suele emplearse en planes de muestreo más complejos. Aunque desde el punto de vista práctico su interés es menor, ya que a igualdad de tamaño de muestra es siempre menos preciso que el m.a.s., el m.a.s. en ocasiones es inviable.

2.2. Selección de una muestra aleatoria simple

Para obtener una m.a.s. a partir de la población anterior basta con tener en cuenta que todas las unidades tienen la misma probabilidad de ser elegidas y que se extraen de la población al azar, de una en una y sin reposición, de modo que en cada extracción todas las unidades que aún no se han extraído tienen la misma probabilidad de ser escogidas para la muestra.

Supongamos que se considera una población de N individuos o unidades que denotaremos por $U_1, \ldots, U_N$ y que van a seleccionarse muestras de tamaño n a partir de la población.

Proposición 2.1. *Todas las m.a.s. de tamaño n a partir de una población de N individuos tienen la misma probabilidad de ser seleccionadas.*

Demostración. Como el número de muestras posibles es $\binom{N}{n}$, vamos a comprobar que en este caso todas las muestras tienen probabilidad $1/\binom{N}{n}$.

Cada muestra podría identificarse con la clase de todas las secuencias ordenadas formadas por las mismas n unidades. Supongamos que se considera la secuencia ordenada $(U_{i_1}, \ldots, U_{i_n})$ donde $i_1, \ldots, i_n \in \{1, \ldots, N\}$, entonces $\Pr(U_{i_1}, \ldots, U_{i_n}) = \Pr(1^{\text{a}} \text{ extr.} = U_{i_1}) \cdot \Pr(2^{\text{a}} \text{ extr.} = U_{i_2} / 1^{\text{a}} \text{ extr.} = U_{i_1}) \cdots \Pr(\text{n}^{\text{a}} \text{ extr.} = U_{i_n} / 1^{\text{a}} \text{ extr.} = U_{i_1}, 2^{\text{a}} \text{ extr.} = U_{i_2}, \ldots, (\text{n-1})^{\text{a}} \text{ extr.} = U_{i_{n-1}})$

$$= \frac{1}{N} \frac{1}{N-1} \cdots \frac{1}{N-(n-1)} = \frac{(N-n)!}{N!} = \frac{1}{V_{N,n}}$$

Como hay un total de $n!$ secuencias ordenadas que corresponden a la misma muestra, la probabilidad de cada muestra aleatoria simple es

$$\Pr(\{U_{i_1}, \ldots, U_{i_n}\}) = n! \frac{(N-n)!}{N!} = \frac{1}{\binom{N}{n}} = \frac{1}{C_{N,n}}$$

En consecuencia, las distintas muestras tendrán todas la misma probabilidad de ser seleccionadas. □

Vamos ahora a analizar la estimación de tres de las características fundamentales en el muestreo.

2.3. Estimación puntual en el m.a.s.

2.3.1. Estimación puntual de la media poblacional

Supongamos que una variable Y toma sobre las unidades $U_1, \ldots, U_N$ de la población los valores (distintos o no) $Y_1, \ldots, Y_N$, respectivamente.

La **MEDIA POBLACIONAL** de Y viene dada por

$$\overline{Y} = \frac{1}{N}\sum_{j=1}^{N} Y_j$$

Supongamos que se extrae una muestra aleatoria simple de tamaño n, s, en la que Y toma los valores $y_{s1}, \ldots, y_{sn}$, entonces la **MEDIA MUESTRAL** se puede expresar por

$$\overline{y}_s = \frac{1}{n}\sum_{i=1}^{n} y_{si}$$

Este valor, para cada muestra concreta, será una estimación de la media poblacional. Vamos a comprobar que cuando se considera el estimador correspondiente sobre el espacio de todas las m.a.s. de tamaño n es un estimador insesgado.

Proposición 2.2. *En el m.a.s. $\overline{y}$ es un estimador insesgado de $\overline{Y}$.*

Demostración. Alternativamente y para que esté bien definido sobre el espacio de las m.a.s., la media muestral puede expresarse como sigue:

$$\overline{y}_s = \frac{1}{n}\sum_{j=1}^{N} Y_j \xi_j(s)$$

donde

$$\xi_j(s) = \begin{cases} 1 & \text{si } U_j \in s \\ 0 & \text{en otro caso} \end{cases}$$

ξ_j es entonces una variable sobre el espacio de las $\binom{N}{n}$ muestras distintas, sobre el que se comporta como una variable con distribución de Bernoulli de parámetro:

$$\pi_j = \Pr(U_j \in s) = \frac{\text{nº de m.a.s. que contienen a } U_j}{\text{nº total de m.a.s.}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

Por lo tanto $\xi_j \sim \mathcal{B}(1, \frac{n}{N})$. Y así,

$$E(\overline{y}) = \sum_s \overline{y}_s \Pr(s) = \sum_s \frac{1}{n}\sum_{j=1}^{N} Y_j \xi_j(s) \Pr(s)$$

$$= \frac{1}{n}\sum_{j=1}^{N} Y_j \sum_s \xi_j(s) \Pr(s) = \frac{1}{n}\sum_{j=1}^{N} Y_j E(\xi_j) = \frac{1}{N}\sum_{j=1}^{N} Y_j = \overline{Y}.$$

En consecuencia, $\overline{y}$ es un estimador insesgado de $\overline{Y}$ en el m.a.s. □

2.3.2. Estimación puntual del total poblacional

Se define el **TOTAL POBLACIONAL** para la variable Y como

$$Y_T = \sum_{j=1}^{N} Y_j = N\overline{Y}$$

Suponiendo N conocido, se puede construir un estimador insesgado para Y_T a partir del estimador insesgado de la media poblacional, y será $N\overline{y}$.

$$E(N\overline{y}) = NE(\overline{y}) = N\overline{Y} = Y_T.$$

2.3.3. Estimación puntual de una proporción poblacional

Supongamos que se considera una población con N unidades y cierta propiedad que los individuos de la población pueden cumplir o no.

Sea R el nº de unidades que satisfacen cierta propiedad concreta. Llamaremos **PROPORCIÓN POBLACIONAL** asociada a esa propiedad al valor:

$$P = \frac{R}{N}$$

Si se define una variable X que para cada unidad U_j de la población toma el valor X_j con:

$$X(U_j) = X_j = \begin{cases} 1 & \text{si } U_j \text{ satisface la propiedad} \\ 0 & \text{en caso contrario} \end{cases}$$

entonces $P = \overline{X}$.

Por lo tanto, P puede contemplarse como una media poblacional y, en consecuencia, la proporción muestral $p = r/n$ donde n es el tamaño muestral y r es el nº de individuos de la muestra que satisfacen esa propiedad, que se puede expresar alternativamente como

$$p = \frac{1}{n}\sum_{j=1}^{N} X_j\xi_j$$

determina un estimador insesgado de P en el m.a.s.

2.3.4. Precisión de los estimadores puntuales

En Técnicas de Muestreo interesa determinar la precisión de un estimador con varios fines:

1. Cuantificar la magnitud de la precisión asociada a un estimador de un parámetro en un procedimiento de muestreo concreto.

2. Comparar para un mismo procedimiento de muestreo y un mismo parámetro, diferentes estimadores en función de la precisión asociada a cada uno de ellos.

3. Comparar dos procedimientos de muestreo diferentes en términos de la precisión que asocia cada uno de ellos a un mismo estimador (o al menos a dos estimadores insesgados) de un mismo parámetro.

4. Determinar tamaños de muestra adecuados para cada procedimiento de muestreo y cada estimador de un parámetro para conseguir un nivel mínimo de precisión con un riesgo máximo admisible.

En la práctica, la precisión va a medirse en función del error cuadrático medio asociado al estimador que se considere. Puesto que para las tres características (media poblacional, total poblacional y proporción poblacional) hemos determinado sendos estimadores insesgados, sus errores cuadrático medios coinciden con sus varianzas.

$$\mathrm{Var}(\overline{y}) = E[(\overline{y})^2] - [E(\overline{y})]^2 = E\left[\left(\frac{1}{n}\sum_{j=1}^{N} Y_j\xi_j\right)^2\right] - \overline{Y}^2$$

$$= \frac{1}{n^2}E\left[\sum_{j=1}^{N} Y_j^2\xi_j^2 + \sum_{j=1}^{N}\sum_{\substack{l=1\\ l\neq j}}^{N} Y_jY_l\xi_j\xi_l\right] - \overline{Y}^2$$

$$= \frac{1}{n^2}\sum_{j=1}^{N} Y_j^2E(\xi_j^2) + \frac{1}{n^2}\sum_{j=1}^{N}\sum_{\substack{l=1\\ l\neq j}}^{N} Y_jY_lE(\xi_j\xi_l) - \overline{Y}^2$$

donde $\xi_j \equiv \mathcal{B}\left(1, \frac{n}{N}\right)$ para $j = 1, \ldots, N$, con lo que $\xi_j = \xi_j^2$ y por lo tanto $E(\xi_j^2) = \frac{n}{N}$.

Por otro lado, si $j \neq l$, $\xi_j\xi_l \equiv \mathcal{B}(1, \pi_{jl})$ con

$$\pi_{jl} = \Pr(\xi_j\xi_l = 1) = \Pr(U_j \text{ y } U_l \text{ intervengan ambas en una misma muestra})$$

$$= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

con lo que $E(\xi_j\xi_l) = \frac{n(n-1)}{N(N-1)}$, de donde,

$$\mathrm{Var}(\overline{y}) = \frac{1}{nN}\sum_{j=1}^{N} Y_j^2 + \frac{(n-1)}{nN(N-1)}\sum_{j=1}^{N}\sum_{\substack{l=1\\ l\neq j}}^{N} Y_jY_l - \overline{Y}^2$$

$$= \frac{N(n-1)}{n(N-1)}\left[\frac{1}{N}\sum_{j=1}^{N} Y_j\right]^2 + \left(\frac{1}{nN} - \frac{(n-1)}{nN(N-1)}\right)\sum_{j=1}^{N} Y_j^2 - \overline{Y}^2$$

$$= \frac{N-n}{n(N-1)}\left[\overline{Y^2} - \overline{Y}^2\right] = \frac{N-n}{n(N-1)}\sigma_Y^2 = \frac{N-n}{nN}S_Y^2 = \frac{1-f}{n}S_Y^2$$

con $f = \frac{n}{N}$ = **fracción muestral** y $1 - f$ = **factor de corrección para poblaciones finitas** y donde S_Y^2 = cuasivarianza poblacional = $\dfrac{1}{N-1}\sum_{j=1}^{N}(Y_j - \overline{Y})^2 = \dfrac{N}{N-1}\sigma_Y^2$.

En resumen,

$$\boxed{\mathrm{Var}(\overline{y}) = \frac{1-f}{n} S_Y^2 \text{ en el m.a.s.}}$$

A partir de ese resultado se concluye sin dificultades que

$$\boxed{\mathrm{Var}(N\overline{y}) = \frac{N^2(1-f)}{n} S_Y^2 \text{ en el m.a.s.}}$$

y que

$$\mathrm{Var}(p) = \frac{1-f}{n} S_X^2 = \frac{(1-f)N}{n(N-1)}\sigma_X^2$$

$$X(U_j) = X_j = \begin{cases} 1 & \text{si } U_j \text{ satisface la propiedad} \\ 0 & \text{en caso contrario} \end{cases}$$

por lo que X es una variable aleatoria Bernoulli de parámetro P = probabilidad de que una unidad satisfaga la propiedad, de donde $\sigma_X^2 = P(1-P)$, de donde

$$\boxed{\mathrm{Var}(p) = \frac{(1-f)N}{n(N-1)} P(1-P) \text{ en el m.a.s.}}$$

2.3.5. Estimación de la precisión

En la práctica suelen desconocerse algunas de las componentes de la varianza de los estimadores, en concreto, las características poblacionales, S_Y^2 y P. Aunque para conseguir los objetivos de los argumentos 2) y 3) para justificar el interés de calcular el error cuadrático medio de los estimadores, no suele ser necesario conocer el valor exacto ni aproximado de estos parámetros desconocidos, sin embargo, para los argumentos 1) y 4) va a ser necesario aproximarlos. En particular, vamos a presentar algunas estimaciones insesgadas de las mismas.

Supongamos que la variable Y toma los valores $Y_j = Y(U_j)$, $j = 1, \ldots, N$ y supongamos que se selecciona una m.a.s. de n individuos de la población y que en esa muestra los valores de la variable Y han sido $y_1, \ldots, y_n$. Llamaremos **cuasivarianza muestral de Y** a

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

Proposición 2.3. *En el m.a.s. el s^2 es un estimador insesgado de S_Y^2.*

Demostración.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\sum_{j=1}^{N}(Y_j - \overline{y})^2\xi_j$$

$$= \frac{1}{n-1}\left[\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2\xi_j+\sum_{j=1}^{N}\left(\overline{Y}-\overline{y}\right)^2\xi_j+2\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)\left(\overline{Y}-\overline{y}\right)\xi_j\right]$$

$$= \frac{1}{n-1}\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2\xi_j+\frac{1}{n-1}\left(\overline{Y}-\overline{y}\right)^2\sum_{j=1}^{N}\xi_j+\frac{2\left(\overline{Y}-\overline{y}\right)}{n-1}\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)\xi_j$$

$$= \frac{1}{n-1}\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2\xi_j+\frac{n}{n-1}\left(\overline{Y}-\overline{y}\right)^2+\frac{2\left(\overline{Y}-\overline{y}\right)}{n-1}\left[\sum_{j=1}^{N}Y_j\xi_j-n\overline{Y}\right]$$

$$= \frac{1}{n-1}\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2\xi_j+\frac{n}{n-1}\left(\overline{Y}-\overline{y}\right)^2-\frac{2n}{n-1}\left(\overline{Y}-\overline{y}\right)^2$$

$$= \frac{1}{n-1}\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2\xi_j-\frac{n}{n-1}\left(\overline{Y}-\overline{y}\right)^2.$$

Como,

$$E\left(\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2\xi_j\right)=\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2E\left(\xi_j\right)=\frac{n}{N}\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2=\frac{n(N-1)}{N}S_Y^2$$

y

$$E\left[n\left(\overline{Y}-\overline{y}\right)^2\right]=nE\left[\left(\overline{y}-E(\overline{y})\right)^2\right]=n\,\mathrm{Var}(\overline{y})=(1-f)S_Y^2$$

se tiene que,

$$E(s^2)=\frac{1}{n-1}\left[\frac{n(N-1)}{N}S_Y^2-\frac{N-n}{N}S_Y^2\right]=S_Y^2,$$

con lo que se concluye el resultado enunciado. □

Así,

$$\boxed{\widehat{\mathrm{Var}(\overline{y})}=\frac{1-f}{n}s^2 \text{ es un estimador insesgado de } \mathrm{Var}(\overline{y}) \text{ en el m.a.s.}}$$

Siguiendo un razonamiento análogo deducimos que,

$$\boxed{\widehat{\mathrm{Var}(N\overline{y})}=\frac{N^2(1-f)}{n}s^2 \text{ es un estimador insesgado de } \mathrm{Var}(N\overline{y}) \text{ en el m.a.s.}}$$

Por otro lado, cuando se considera la estimación de la proporción poblacional P a partir de la muestral p, será:

$$s_X^2=\frac{1}{n-1}\sum_{j=1}^{N}(X_j-p)^2\xi_j=\frac{1}{n-1}\left[\sum_{j=1}^{N}X_j^2\xi_j+\sum_{j=1}^{N}p^2\xi_j-2p\sum_{j=1}^{N}X_j\xi_j\right]$$

$$=\frac{1}{n-1}\left[np+np^2-2np^2\right]=\frac{n}{n-1}p(1-p)$$

y, en consecuencia:

$$\boxed{\widehat{\mathrm{Var}(p)} = \frac{1-f}{n-1}p(1-p) \text{ es un estimador insesgado de } \mathrm{Var}(p) \text{ en el m.a.s.}}$$

La precisión de un estimador suele expresarse en términos de su **error de muestreo**, es decir, de la raíz cuadrada positiva de su error cuadrático medio, ya que de esta forma la precisión se mide en las mismas unidades que la variable.

2.4. Estimación por intervalos en el m.a.s.

2.4.1. Estimación por intervalo de la media

Cuando se considera la estimación puntual, el error subyacente a la inferencia se valora en términos de medidas del sesgo, error típico, error cuadrático medio, etc., pero no puede valorarse en términos de la probabilidad de que dicho error sea más o menos elevado.

Para hallar estimaciones cuya probabilidad de error máximo pueda controlarse, y para las que pueda indicarse el grado de seguridad y el grado de precisión en términos de probabilidades con interpretación intuitiva útil, vamos a considerar la **ESTIMACIÓN POR INTERVALOS DE CONFIANZA**. A menudo, muchos de los estimadores puntuales se "corrigen" por el error de muestreo para dar lugar a los estimadores por intervalo.

Aproximación basada en la desigualdad de Tchebychev

Una primera aproximación al problema, que proporciona intervalos "por exceso", pero es aplicable para cualquier tamaño muestral, es la basada en la **desigualdad de Tchebychev**, según la cual para cualquier variable aleatoria Z:

$$\Pr(|Z - E(Z)| \leq \varepsilon) \geq 1 - \frac{\mathrm{Var}(Z)}{\varepsilon^2} \qquad \forall\, \varepsilon > 0$$

Supongamos que queremos construir un intervalo de confianza para el parámetro $\overline{Y}$ a partir de una m.a.s. de tamaño n de la población y con coeficiente de confianza $1-\alpha$ ($\alpha \in [0,1]$), entonces basta con tomar en la desigualdad de Tchebychev $\alpha = \dfrac{\mathrm{Var}(Z)}{\varepsilon^2}$ con lo que $\varepsilon = \sqrt{\dfrac{\mathrm{Var}(Z)}{\alpha}}$, y se tendría que

$$\Pr\left(|\overline{y} - \overline{Y}| \leq \sqrt{\frac{\mathrm{Var}(\overline{y})}{\alpha}}\right) \geq 1 - \alpha.$$

Por lo tanto, un intervalo de confianza para la media poblacional con coeficiente de confianza $1-\alpha$ vendrá dada por:

$$\left(\overline{y} - \sqrt{\frac{\mathrm{Var}(\overline{y})}{\alpha}}, \overline{y} + \sqrt{\frac{\mathrm{Var}(\overline{y})}{\alpha}}\right).$$

En el muestreo aleatorio simple, será

$$I.C(\overline{Y})_{1-\alpha} = \left(\overline{y} - \sqrt{\frac{1-f}{n\alpha}}S_Y, \overline{y} + \sqrt{\frac{1-f}{n\alpha}}S_Y\right)$$

En la práctica, difícilmente se conoce el valor de S_Y^2, por tratarse de una característica poblacional, por lo que suele reemplazarse por la cuasivarianza de Y en la muestra, s^2. Como en la desigualdad de Tchebychev se garantiza habitualmente una probabilidad muy superior a la que acota superiormente, la proporción de intervalos de confianza que contendrían al verdadero valor del parámetro suele ser muy superior a $1-\alpha$ y la proporción de los intervalos que contienen al verdadero valor del parámetro cuando se reemplaza S_Y^2 por s^2 sigue siendo habitualmente superior a $1-\alpha$.

Como a menudo los intervalos proporcionados por la desigualdad de Tchebychev son muy amplios, es deseable proponer otros intervalos mediante alguna aproximación diferente. Si la distribución de $\overline{y}$ se conociera con exactitud, al menos por lo que se refiere a su clase paramétrica, sería posible buscar tal aproximación.

Aproximación normal

Al realizarse un muestreo sin reposición, las observaciones muestrales y_i están idénticamente distribuidas pero no son independientes, por lo que en principio no puede aplicarse directamente el Teorema Central del Límite, y aún en el caso en que la variable Y tuviera distribución normal tampoco podría aplicarse la reproductividad de la normal. Hájek dio una condición necesaria y suficiente bajo la cual la distribución de $\overline{y}$ es asintóticamente normal:

Proposición 2.4. *Se dispone de una sucesión de poblaciones de tamaños $N_1, \ldots, N_m$ con*

$$\lim_{m\to\infty} N_m = \infty.$$

En cada una de esas poblaciones la variable Y toma los valores $Y_{m1}, \ldots, Y_{mN_m}$ con lo que la media de Y en la m-ésima población es

$$\overline{Y}_m = \frac{1}{N_m}\sum_{j=1}^{N_m} Y_{mj}$$

y la cuasivarianza de Y en la m-ésima población viene dada por

$$S_m^2 = \frac{1}{N_m-1}\sum_{j=1}^{N_m}(Y_{mj}-\overline{Y}_m)^2$$

Supongamos que en la m-ésima población se extrae una muestra aleatoria sin reposición de tamaño n_m con

$$\lim_{m\to\infty} n_m = \infty \qquad y \qquad \lim_{m\to\infty}(N_m - n_m) = \infty$$

Si además se satisface la condición de Lindeberg:

$$\lim_{m\to\infty} \frac{\sum_{C_{mk}}(Y_{mj}-\overline{Y}_m)^2}{(N_m-1)S_m^2} = 0 \qquad \forall\, k>0$$

donde $C_{mk} = \{j = 1, \ldots, N \,/\, |Y_{mj}-\overline{Y}_m| > k\sqrt{n_m(1-f_m)}S_m\}$ y $f_m = n_m/N_m$.

Entonces, ésta es una condición necesaria y suficiente para garantizar que $\overline{y}_m$ sea asintóticamente normal con media $\overline{Y}_m$ y varianza $\frac{1-f_m}{n_m}S_m^2$, es decir,

$$\overline{y}_m \sim N\left(\overline{Y}_m, \sqrt{\frac{1-f_m}{n_m}}S_m\right)$$

La complejidad práctica de este resultado estriba en la comprobación de la condición de Lindeberg y en la determinación de los tamaños de muestra que permiten que la aproximación asintótica sea adecuada. La respuesta a esa cuestión no es sencilla, puesto que las distribuciones poblacionales son muy diversas y los tamaños mínimos variarían de unas a otras. Con frecuencia, el "número mágico" $n = 30$ no basta en el muestreo de las poblaciones finitas. Si la distribución se asemeja a la normal, probablemente sea seguro utilizar el teorema con un tamaño de muestra de 50.

A través de técnicas de simulación se ha comprobado que cuando la población es tal que la variable Y tiene distribución marcadamente asimétrica por la derecha (como por ejemplo las distribuciones de Pareto, logaritmo normal, Gamma,...) una regla conservadora que suele ser útil es tomar $n \geq 25\gamma_1^2$ con $\gamma_1 =$ coeficiente de asimetría de Fisher,

$$\gamma_1 = \frac{1}{N}\sum_{j=1}^{N}\frac{(Y_j - \overline{Y})^3}{S^3}$$

o más concretamente un estimador de este coeficiente.

Para muestras pequeñas con distribuciones muy asimétricas, son necesarios métodos especiales. Por ejemplo, cuando la población es marcadamente asimétrica y aparecen unos cuantos valores extremos, estos valores tendrán una influencia muy grande sobre la media y más aún sobre la varianza. Para tratar de disminuir esta varianza, un procedimiento posible es recurrir al muestreo estratificado incluyendo los valores más extremos en diferentes estratos.

Después de las consideraciones anteriores, cuando sea posible aplicar la aproximación normal, (convenimos para $n \geq 100$ ó $n \geq 25\gamma_1^2$)

$$\left(\overline{y} - z_{1-\alpha/2}\sqrt{\frac{1-f}{n}}S, \overline{y} + z_{1-\alpha/2}\sqrt{\frac{1-f}{n}}S\right)$$

constituye un intervalo de confianza para la media poblacional con coeficiente aproximadamente $1-\alpha$, donde $z_{1-\alpha/2}$ es el percentil $1-\alpha/2$ de la distribución normal estándar.

Por lo general, el valor de S^2 se desconoce, de modo que hay que estimarlo por el correspondiente estimador insesgado s^2. Si el tamaño de la muestra es menor de 60, $z_{1-\alpha/2}$ suele reemplazarse por el valor $t_{1-\alpha/2}$ de la distribución t de Student con $n-1$ grados de libertad.

2.4.2. Estimación por intervalo del total

El estudio de límites de confianza para Y_T es inmediato a partir del de $\overline{Y}$. Basta con considerar los intervalos correspondientes a la media y multiplicar los dos extremos por el tamaño poblacional, N.

2.4.3. Estimación por intervalo de la proporción

La estimación por intervalo de P puede considerarse como un caso particular del de la estimación por intervalo de $\overline{Y}$. Sin embargo, en este caso se sabe que r se distribuye según una hipergeométrica de parámetros N, R y n. Esto se podría utilizar para construir intervalos de confianza exactos para P, pero como en la práctica se requiere una gran cantidad de cálculos y el manejo de tablas de la distribución hipergeométrica, usualmente se recurre a procedimientos aproximados.

Aproximación basada en la desigualdad de Tchebychev

Basándonos en la desigualdad de Tchebychev, sabemos que

$$\Pr\left(|p-P| \leq \sqrt{\frac{\mathrm{Var}(p)}{\alpha}}\right) \geq 1-\alpha$$

de donde

$$\Pr\left(|p-P| \leq \sqrt{\frac{(1-f)NP(1-P)}{n(N-1)\alpha}}\right) \geq 1-\alpha.$$

Como

$$|p-P| \leq \sqrt{\frac{(1-f)NP(1-P)}{n(N-1)\alpha}} \iff (p-P)^2 \leq \frac{(1-f)NP(1-P)}{n(N-1)\alpha}$$

$$\iff n(N-1)\alpha p^2 + (N-1)n\alpha P^2 - 2pP(N-1)n\alpha \leq N(1-f)P - N(1-f)P^2.$$

Agrupando convenientemente los términos en P, llegamos a una inecuación de segundo grado de la forma

$$aP^2 + bP + c \leq 0 \qquad (a = (N-1)n\alpha + N(1-f) > 0).$$

Si p_1 y p_2 son las dos raíces de la ecuación $aP^2 + bP + c = 0$ con $p_1 \leq p_2$ se tiene que

$$|p-P| \leq \sqrt{\frac{(1-f)NP(1-P)}{n(N-1)\alpha}} \iff a(P-p_1)(P-p_2) \leq 0 \iff p_1 \leq P \leq p_2.$$

Con lo que $[p_1, p_2]$ es un intervalo de confianza para P con coeficiente de confianza $1-\alpha$ en el m.a.s. (habitualmente muy conservador).

Aproximación normal

Si n es relativamente pequeño con respecto a R y $N-R$, r sigue esencialmente una distribución $\mathcal{B}(n, P)$. Se podría utilizar la distribución binomial para construir intervalos de confianza para P, pero de nuevo nos encontramos con dificultades en los cálculos. En la mayoría de las aplicaciones se da un paso más y se aproxima la distribución binomial por la distribución normal. Entonces,

$$p \to \mathcal{N}\left(P, \sqrt{\frac{(1-f)P(1-P)}{n}}\right).$$

En dicha aproximación se ha tenido en cuenta la falta de reemplazamiento, incorporando el factor de corrección para poblaciones finitas en $\mathrm{Var}(p)$. También se ha prescindido del término $\dfrac{N}{N-1}$, ya que los tamaños de población donde se aplica la aproximación normal lo justifican.

Resumiendo, la aproximación normal es razonable cuando el tamaño poblacional N es suficientemente grande y el tamaño muestral n grande pero de forma que la fracción muestral $f = n/N$ no resulte muy elevada (tomaremos $N \geq 60$ y $n/N \leq 0{,}1$), y se cumple la condición

$$\text{mín}(nP, n(1-P)) \simeq \text{mín}(np, n(1-p)) \geq 30.$$

A partir de la aproximación normal vamos a decidir dos procedimientos. El primero es el que proporciona mejores aproximaciones.

Primer procedimiento

Para hallar un intervalo de confianza con coeficiente $1-\alpha$ tomamos

$$\Pr\left(|p-P| \leq z_{1-\alpha/2}\sqrt{\frac{(1-f)P(1-P)}{n}}\right) \simeq 1-\alpha.$$

Como en la aproximación basada en la desigualdad de Tchebychev,

$$|p-P| \leq z_{1-\alpha/2}\sqrt{\frac{(1-f)P(1-P)}{n}} \iff (p-P)^2 \leq z^2_{1-\alpha/2}\frac{(1-f)P(1-P)}{n}.$$

De forma que un intervalo de confianza para P es el que determinan las dos raíces de la ecuación

$$[1+z^2_{\alpha/2}(1-f)/n]P^2 - [2p+z^2_{\alpha/2}(1-f)/n]P + p^2 = 0.$$

Segundo procedimiento

Cuando n es suficientemente grande, en lugar de despejar P como antes, se reemplaza $\mathrm{Var}(p)$ por su estimación insesgada $\dfrac{(1-f)}{n-1}p(1-p)$, con lo que obtenemos como intervalo:

$$\left[p - t_{1-\alpha/2}\sqrt{\frac{(1-f)p(1-p)}{n-1}}, p + t_{1-\alpha/2}\sqrt{\frac{(1-f)p(1-p)}{n-1}}\right].$$

Se ha comprobado que los intervalos resultantes son a veces poco amplios, de modo que la proporción de muestras que conducen a intervalos de confianza que contienen a P suele ser inferior a $1-\alpha$. Para evitar estos inconvenientes, suele recurrirse a la llamada corrección por continuidad que trata de corregir la "no continuidad" de la distribución verdadera de p y que consiste simplemente en añadirle en la amplitud por cada lado $\frac{1}{2n}$.

Según esta corrección, el intervalo sería:

$$\left[p - t_{1-\alpha/2}\sqrt{\frac{(1-f)p(1-p)}{n-1}} - \frac{1}{2n}, p + t_{1-\alpha/2}\sqrt{\frac{(1-f)p(1-p)}{n-1}} + \frac{1}{2n}\right]$$

A través de técnicas de simulación se han determinado reglas de trabajo para decidir cuándo la aproximación normal anterior puede ser usada (Ver Cochran):

Cuadro 2.1: VALORES MÁS PEQUEÑOS DE np PARA USO DE LA APROXIMACIÓN NORMAL

p	np = nº observado en la clase más pequeña	n = tamaño muestral
0'5	15	30
0'4	20	50
0'3	24	80
0'2	40	200
0'1	60	600
0'05	70	1400
~ 0	80	∞

2.5. Elección de tamaños de muestra en el m.a.s.

Es evidente, a partir de los estudios precedentes, que cuanto mayor es el tamaño de la muestra, menor es la varianza de los estimadores y en consecuencia, la estimación resulta más precisa. Sin embargo, un aumento notable del tamaño muestral supone un aumento muy grande en el coste económico, de tiempo, etc...

En la determinación del tamaño de muestra idóneo, los pasos principales que van a seguirse son los siguientes:

1. Debe haber una indicación relativa a lo que se espera de la muestra. Esta indicación puede estar en términos de los límites de error deseados o en términos de alguna decisión que se va a hacer o acción que se va a tomar cuando los resultados de la muestra sean conocidos. En la práctica este paso lo deberían desarrollar personas expertas en el tema en el que se aplique, convenientemente asesoradas por expertos estadísticos.

2. Establecer una ecuación o inecuación que conecte el tamaño muestral con la precisión deseada. La ecuación variará con la especificación de la precisión y con el procedimiento de muestreo considerado. Una de las ventajas de los muestreos probabilísticos, es que permiten construir esa ecuación. Habitualmente esta ecuación o inecuación suele ser una condición suficiente aunque no equivalente del objetivo que nosotros perseguimos, con lo que el tamaño muestral resultante no será realmente el menor posible para garantizar las condiciones iniciales sino uno algo superior a él.

3. En muchos casos, en el paso 2 aparece involucrado algún valor poblacional desconocido, habitualmente estos valores se reemplazan por alguna estimación y es importante tener en cuenta que estas estimaciones no son nuestro objetivo sino que simplemente vamos a emplearlas como soporte para nuestro objetivo prioritario que es la estimación de otro parámetro poblacional.

4. Debe verificarse si el tamaño muestral resultante es consistente con los recursos disponibles para tomar la muestra. Esto demanda una estimación del costo, trabajo, tiempo y materiales requeridos para obtener el tamaño de muestra propuesto. En el caso de que n deba ser reducida drásticamente, suelen debilitarse las condiciones sobre la precisión fijada.

El error de muestreo puede venir dado en términos absolutos, en términos relativos o sujeto adicionalmente a un coeficiente de confianza dado. A continuación vamos a estudiar el procedimiento a seguir en el caso de estimar el parámetro que interesa, θ, con un *error máximo admisible o tolerancia*, e_α, (máxima diferencia entre la estimación y el valor verdadero, que estamos dispuestos a asumir), y un *coeficiente de confianza*, $1-\alpha$. El objetivo, por tanto, va a ser determinar el menor n tal que

$$\Pr(|\hat{\theta}_n - \theta| \leq e_\alpha) \geq 1-\alpha$$

siendo $\hat{\theta}_n$ un estimador de θ a partir de una m.a.s. de tamaño n.

Consideremos la determinación del tamaño de muestra adecuado en la estimación de medias, totales y proporciones poblacionales, cuando se considera el muestreo aleatorio simple y los estimadores insesgados que hemos visto anteriormente.

MEDIA POBLACIONAL

Supongamos que se quiere estimar la media poblacional $\overline{Y}$ mediante la media de una muestra aleatoria simple, $\overline{y}$, suponiendo un nivel aceptable $1-\alpha$ para la probabilidad de que $|\overline{y}-\overline{Y}| \leq e_\alpha$ (precisión deseada). Si no se hace ninguna consideración expresa sobre los costes, se supone que éstos son directamente proporcionales al tamaño de la muestra, por lo que se elegirá el mínimo n que satisfaga la condición:

$$\Pr(|\overline{y}-\overline{Y}| \leq e_\alpha) \geq 1-\alpha.$$

Supongamos que n va a ser suficientemente grande como para poder aplicar la aproximación normal que admite que $\overline{y} \sim \mathcal{N}\left(\overline{Y}, \sqrt{\mathrm{Var}(\overline{y})}\right)$. Se verificará entonces que

$$\Pr\left(|\overline{y}-\overline{Y}| \leq z_{1-\alpha/2}\sqrt{\mathrm{Var}(\overline{y})}\right) \simeq 1-\alpha$$

con lo cual si el tamaño muestral n se toma de modo que sea el menor entero tal que $z_{1-\alpha/2}\sqrt{\mathrm{Var}(\overline{y})} \leq e_\alpha$ se satisfará aproximadamente el objetivo deseado.

Esto significa que,

$$\left(\frac{1}{n}-\frac{1}{N}\right)S_Y^2 \leq \frac{e_\alpha^2}{z_{1-\alpha/2}^2} \iff n \geq \left(\frac{1}{N}+\frac{e_\alpha^2}{z_{1-\alpha/2}^2 S_Y^2}\right)^{-1} \iff n \geq \frac{\left(\frac{z_{1-\alpha/2}S_Y}{e_\alpha}\right)^2}{1+\frac{1}{N}\left(\frac{z_{1-\alpha/2}S_Y}{e_\alpha}\right)^2}$$

de modo que si S_Y^2 es conocido, podemos tomar

$$n = \left]\frac{\left(\frac{z_{1-\alpha/2}S_Y}{e_\alpha}\right)^2}{1+\frac{1}{N}\left(\frac{z_{1-\alpha/2}S_Y}{e_\alpha}\right)^2}\right[$$

con $]$ = mayor parte entera (número entero más próximo por exceso).

Una primera aproximación al tamaño de muestra requerido vendrá dado por

$$n_0 = \left(\frac{z_{1-\alpha/2}S_Y}{e_\alpha}\right)^2 = \frac{z_{1-\alpha/2}^2 S_Y^2}{e_\alpha^2}.$$

En la práctica se calcula primero n_0. Si n_0/N es despreciable, n_0 es una aproximación adecuada de n. Si no es así, n se obtiene como sigue

$$n = \frac{n_0}{1+\frac{n_0}{N}}$$

Esto presupone, sin embargo, que S_Y^2 es conocida. Esta posibilidad es remota en la práctica y en tales circunstancias suele recurrirse a una de las cuatro vias siguientes:

- Consideración de un estudio piloto, que proporciona a menudo informaciones adicionales (marcos de muestreo posibles, dificultades que pueden presentarse en el muestreo de la población considerada, etc.) Si la muestra piloto es una m.a.s. de la población, S_Y^2 puede estimarse mediante s^2. Sin embargo, en ocasiones la muestra piloto se extrae de una parte de la población y el estimador resulta a menudo bastante sesgado. Así, por ejemplo, es relativamente común reducir el muestreo piloto a unos pocos conglomerados y la s^2 suele infravalorar S_Y^2.

- Consideración de estudios previos realizados sobre características y poblaciones similares cuya información puede aprovecharse para estimar S_Y^2 aunque a veces sea necesario algún ajuste, debido generalmente a los cambios en el tiempo.

- Cuando la característica corresponde a una variable cuya distribución se ajusta bien a un modelo conocido, a veces es posible aprovechar esta información. Por ejemplo, si la distribución de la característica considerada fuera de tipo Poisson, se sabe que la media y la varianza tienen el mismo valor y puede estimarse el valor de la varianza mediante el valor de la media siempre que se disponga alguna estimación previa de ésta.

- Se puede considerar una submuestra inicial de tamaño n_1, no muy elevado, de forma que se seleccione con una m.a.s. Se estima sobre la base de esta muestra el valor de S_Y^2 mediante la cuasivarianza de esa submuestra, s_1^2, y se reemplaza el valor poblacional por su estimación, obteniendo un valor de n

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

donde

$$n_0 = \frac{z_{1-\alpha/2}^2 s_1^2}{e_\alpha^2}\left(1 + \frac{2}{n_1}\right)$$

y si $n > n_1$ se extrae otra muestra de tamaño $n - n_1$ hasta completar una muestra de tamaño n.

Este es es el procedimiento más habitual cuando se recurre a la aproximación normal.

Cuando el tamaño de muestra obtenido no verifica las condiciones que garantizan que la distribución de $\overline{y}$ es aproximadamente normal (consideraremos $n \geq 100$) se puede recurrir al tamaño muestral que obtendremos para el mismo problema en el muestreo aleatorio con reposición o bien a la desigualdad de Tchebychev, en virtud de la cual

$$\Pr(|\overline{y} - \overline{Y}| \leq e_\alpha) \geq 1 - \frac{\mathrm{Var}(\overline{y})}{e_\alpha^2}$$

y en consecuencia, una condición suficiente para garantizar la condición anterior es que se cumpla que

$$1 - \frac{\mathrm{Var}(\overline{y})}{e_\alpha^2} \geq 1 - \alpha \iff \mathrm{Var}(\overline{y}) \leq \alpha e_\alpha^2 \iff \frac{1}{n} - \frac{1}{N} \leq \frac{\alpha e_\alpha^2}{S_Y^2}$$

de modo que si S_Y^2 es conocido, podemos tomar n tal que

$$n = \frac{n_0}{1 + \frac{n_0}{N}}\Bigg]$$

donde $n_0 = \dfrac{S_Y^2}{\alpha e_\alpha^2}$.

TOTAL POBLACIONAL

Puede aplicarse lo anterior con la modificación pertinente.

PROPORCIÓN POBLACIONAL

Nuestro objetivo va a ser determinar el menor n tal que

$$\Pr(|p - P| \leq e_\alpha) \geq 1 - \alpha$$

Siempre que se suponga válida la consideración de la distribución normal para modelizar asintóticamente la distribución de p (ver sección anterior), se tiene que,

$$\Pr\left(|p - P| \leq z_{1-\alpha/2}\sqrt{\frac{(1-f)NP(1-P)}{n(N-1)}}\right) \simeq 1 - \alpha$$

Por lo tanto, la fórmula que conecta n con el grado de precisión deseado viene dada por

$$z_{1-\alpha/2}\sqrt{\frac{(1-f)NP(1-P)}{n(N-1)}} \leq e_\alpha \iff \left(\frac{1}{n} - \frac{1}{N}\right)\frac{N}{N-1}P(1-P) \leq \frac{e_\alpha^2}{z_{1-\alpha/2}^2}$$

$$\iff n \geq \left(\frac{1}{N} + \frac{(N-1)e_\alpha^2}{z_{1-\alpha/2}^2 NP(1-P)}\right)^{-1}$$

Se tiene

$$n = \left\lceil \frac{\dfrac{z_{1-\alpha/2}^2 P(1-P)}{e_\alpha^2}}{1 + \dfrac{1}{N}\left(\dfrac{z_{1-\alpha/2}^2 P(1-P)}{e_\alpha^2} - 1\right)} \right\rceil$$

Como antes, una primera aproximación al tamaño muestral requerido viene dada por

$$n_0 = \frac{z_{1-\alpha/2}^2 P(1-P)}{e_\alpha^2}$$

Esto es adecuado a menos que n_0/N sea apreciable, en cuyo caso calculamos n como

$$n = \frac{n_0}{1 + \dfrac{(n_0 - 1)}{N}}$$

No obstante, el valor de P es desconocido y deberemos valorar en alguna forma el producto $P(1-P)$. Para estimar este producto y teniendo en cuenta que esta estimación es un objetivo secundario, conviene estimarlo con la mayor precisión posible o al menos sobrevalorarlo.

Las técnicas usuales para estimar este producto cuando es un objetivo secundario son las siguientes:

- Estimar el producto $P(1-P)$ por el máximo valor que puede tomar este producto según la información disponible.

  - Si por ejemplo se desconoce el rango de posibles valores de P, un posible camino es adoptar una postura conservadora estimándolo por su valor máximo

    $$\widehat{P(1-P)} = \max_{P\in[0,1]} P(1-P) = \frac{1}{4}$$

    ya que $P(1-P)$ alcanza máximo absoluto en el punto $P = 1/2$.
  - Si se sabe que P toma valores en un determinado conjunto $\mathcal{P}$ contenido en $[0,1]$,

    $$\widehat{P(1-P)} = \max_{P\in\mathcal{P}} P(1-P)$$

    Así, si $\mathcal{P} = [P_1, P_2] \subset [0,1]$

    $$\begin{array}{lll} \text{si} \quad 0{,}5 \in [P_1,P_2] & \Rightarrow & \widehat{P(1-P)} = \frac{1}{4} \\ \text{si} \quad P_2 < 0{,}5 & \Rightarrow & \widehat{P(1-P)} = P_2(1-P_2) \\ \text{si} \quad P_1 > 0{,}5 & \Rightarrow & \widehat{P(1-P)} = P_1(1-P_1) \end{array}$$

- Esta segunda técnica consiste en estimar el producto $P(1-P)$ a partir del valor $p_1(1-p_1)$ donde p_1 es la proporción muestral derivada de una muestra piloto u obtenida en una primera muestra de tamaño n_1 extraída de la población según un muestreo aleatorio simple, de modo que completamos la muestra inicial con otra muestra de tamaño $n - n_1$.

Cuando la aproximación normal no sea posible (consideraremos $N \geq 60$ y $n/N \leq 0{,}1$), se puede recurrir a la desigualdad de Tchebychev, según la cual la condición $\Pr(|p-P| \leq e_\alpha) \geq 1-\alpha$ queda garantizada cuando:

$$\frac{(N-n)P(1-P)}{n(N-1)} \leq \alpha e_\alpha^2$$

de forma que, si $n_0 = \dfrac{P(1-P)}{\alpha e_\alpha^2}$, se llega a

$$n = \frac{n_0}{1 + \dfrac{(n_0-1)}{N}}$$

Muestreo aleatorio con reposición

2.6. Selección de una muestra aleatoria con reposición

Supongamos que se considera una población de N individuos y que van a seleccionarse muestras de tamaño n a partir de la población. El **muestreo aleatorio con reposición** consiste en extraer las unidades de la muestra al azar, de una en una y con reposición, de modo que en cada extracción todas las unidades (extraídas o no anteriormente) tienen la misma probabilidad de ser escogidas para la muestra.

Al igual que en el m.a.s., al seleccionarse al azar las unidades, es decir, al tener cada unidad probabilidad $1/N$ de ser la extraída en una extracción concreta, la probabilidad de una secuencia ordenada concreta en el muestreo aleatorio con reposición de tamaño n será:

$$\Pr(U_{i_1},\ldots,U_{i_n}) = \frac{1}{N}\frac{1}{N}\cdots\frac{1}{N} = \frac{1}{N^n} = \frac{1}{VR_{N,n}}$$

y $VR_{N,n}$ es precisamente el número de secuencias ordenadas diferentes de n elementos de la población.

A diferencia del m.a.s., en el m.a. con reposición las probabilidades de las distintas muestras de un mismo tamaño no tienen por qué coincidir.

Si, como hicimos en el muestreo aleatorio simple, entendemos por MUESTRA de tamaño n el conjunto de secuencias ordenadas que tienen en común las unidades que aparecen en ellas y el número de veces que aparece cada unidad, de forma que dos muestras serán distintas si alguna de sus unidades es distinta o aparece un número diferente de veces en cada muestra, se tiene que el número de muestras posibles será $CR_{N,n} = \binom{N+n-1}{n}$. No obstante, cada una de esas muestras no tiene necesariamente la misma que la de las demás muestras. Así, por ejemplo,

$$\Pr(\{U_{i_1},\ldots,U_{i_1}\}) = \frac{1}{VR_{N,n}} = \frac{1}{N^n} \neq \Pr(\{U_{i_1},\ldots,U_{i_n}\}) = \frac{n!}{VR_{N,n}} = \frac{n!}{N^n}$$

2.7. Estimación puntual en el m.a.c.r.

Para estimar la media poblacional, $\overline{Y}$, a partir de una muestra aleatoria con reposición de tamaño n, consideramos como estimador la media muestral, $\overline{y}$.

Este estimador asociaría a cada m.a.c.r., s,

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{j=1}^{N} Y_j t_j(s)$$

donde la variable t_j cuantifica el "número de veces que la unidad U_j aparece en la muestra" y está definida sobre el espacio de las $\binom{N+n-1}{n}$ muestras distintas, espacio sobre el que se distribuye según una binomial $\mathcal{B}(n, 1/N)$, ya que $\frac{1}{N}$ = probabilidad de elegir U_j en una extracción cualquiera ($j = 1,\ldots,N$).

Se puede comprobar que cuando se considera el estimador correspondiente sobre el espacio de todas las m.a.c.r. de tamaño n es un estimador insesgado, es decir, $E(\overline{y}) = \overline{Y}$.

Razonando como en el m.a.s., concluimos de forma inmediata que:

- $y_T = N\overline{y}$ es un estimador insesgado de Y_T en el m.a.c.r.
- p es un estimador insesgado de P en el m.a.c.r.

2.7.1. Precisión de los estimadores puntuales

Al ser $t_j \sim \mathcal{B}(n, 1/N)$ se verifica que $\text{Var}(t_j) = n\frac{1}{N}\left(1-\frac{1}{N}\right) = \frac{n(N-1)}{N^2}$. Por otro lado, la variable N-dimensional $(t_1, \ldots, t_N) \sim \mathcal{M}(n, \frac{1}{N}, \ldots, \frac{1}{N})$, de modo que $\text{Cov}(t_j, t_l) = \frac{-n}{N^2}$ para $j \neq l$ $(j, l \in \{1, \ldots, N\})$. Por lo tanto,

$$\text{Var}(\overline{y}) = \frac{\sigma_Y^2}{n} = \frac{(N-1)}{nN} S_Y^2 = \frac{S_Y^2}{n}\left(1-\frac{1}{N}\right).$$

Como $1-\frac{1}{N} > 1-\frac{n}{N} = 1-f$, $\text{Var}(\overline{y})$ es inferior en el m.a.s. que en el m.a.c.r., y la diferencia se hace más patente cuanto mayor es el tamaño de la muestra. El muestreo aleatorio simple es más preciso que el muestreo aleatorio con reposición, y de hecho para conseguir tanta precisión en el segundo como en el primero con una muestra de tamaño n se requiere una muestra de tamaño

$$m = \left\lceil \frac{n(N-1)}{N-n} \right\rceil$$

En general, puede concluirse que

- $\text{Var}(N\overline{y}) = \frac{N^2\sigma_Y^2}{n}$ en el m.a.c.r.
- $\text{Var}(p) = \frac{P(1-P)}{n}$ en el m.a.c.r.

2.7.2. Estimación de la precisión de los estimadores

El interés de conocer $\text{Var}(\overline{y})$, $\text{Var}(N\overline{y})$ y $\text{Var}(p)$ en el muestreo aleatorio con reposición es el mismo que tenía en el muestreo aleatorio simple. Por ello, es útil estimar estas varianzas cuando se ignoran algunas características poblacionales que figuran en ellas. Vamos a comprobar que la cuasivarianza muestral, s^2, es un estimador insesgado de la varianza poblacional, σ_Y^2.

En efecto,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{n-1}\sum_{j=1}^{N}(Y_j-\overline{y})^2 t_j = \frac{1}{n-1}\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2 t_j - \frac{n}{n-1}\left(\overline{Y}-\overline{y}\right)^2$$

(sin más que razonar como en el m.a.s.) y como

$$E\left(\sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2 t_j\right) = \sum_{j=1}^{N}\left(Y_j-\overline{Y}\right)^2 E(t_j) = n\sigma_Y^2$$

y

$$E\left[n\left(\overline{Y}-\overline{y}\right)^2\right]=n\,\mathrm{Var}(\overline{y})=\sigma_Y^2$$

se tiene que,

$$E(s^2)=\frac{1}{n-1}\left[n\sigma_Y^2-\sigma_Y^2\right]=\sigma_Y^2.$$

Por otro lado, para la variable $X\sim\mathcal{B}(1,P)$, con $P=$ proporción poblacional, se cumple que:

$$s_X^2=\frac{1}{n-1}\sum_{j=1}^{N}(X_j-p)^2t_j=\frac{n}{n-1}p(1-p)$$

En consecuencia, concluimos que:

- $\dfrac{s^2}{n}=\dfrac{1}{n(n-1)}\displaystyle\sum_{i=1}^{n}(y_i-\overline{y})^2$ es un estimador insesgado de $\mathrm{Var}(\overline{y})$ en el m.a.c.r.
- $\dfrac{N^2s^2}{n}=\dfrac{N^2}{n(n-1)}\displaystyle\sum_{i=1}^{n}(y_i-\overline{y})^2$ es un estimador insesgado de $\mathrm{Var}(N\overline{y})$ en el m.a.c.r.
- $\dfrac{p(1-p)}{n-1}$ es un estimador insesgado de $\mathrm{Var}(p)$ en el m.a.c.r.

2.8. Estimación por intervalos en el m.a.c.r.

2.8.1. Estimación por intervalo de la media

En la estimación por intervalo de la media pueden considerarse la aproximación basada en la desigualdad de Tchebychev y la aproximación normal, si bien en ocasiones la naturaleza de la variable permite establecer la distribución de la misma y aplicar resultados relativos a la media de una muestra de observaciones independientes (reproductividad, etc.).

Aproximación basada en la desigualdad de Tchebychev

Parte de la desigualdad

$$\Pr\left(|\overline{y}-\overline{Y}|\le\sqrt{\frac{\sigma_Y^2}{n\alpha}}\right)\ge 1-\alpha$$

Y suponiendo conocida de modo aproximado σ_Y^2, si despejamos la media poblacional, entonces

$$\left[\overline{y}-\frac{\sigma_Y}{\sqrt{n\alpha}},\overline{y}+\frac{\sigma_Y}{\sqrt{n\alpha}}\right]$$

es un intervalo de confianza para $\overline{Y}$ con coeficiente $1-\alpha$ en el m.a.c.r.

Si se desconoce el valor de σ_Y^2, y teniendo en cuenta que la desigualdad de Tchebychev proporciona intervalos habitualmente muy conservadores, el reemplazo de σ_Y^2 por su estimación insesgada s^2

garantiza en la mayoría de los casos que el intervalo correspondiente sigue teniendo un coeficiente de confianza $1-\alpha$.

Aproximación normal basada en el TLC

Al ser ahora las observaciones muestrales independientes, sí podrá aplicarse el Teorema Central del Límite. Si se supone que el tamaño de muestra es suficientemente grande ($n \geq 30$) como para garantizar que la aplicación del TLC es adecuada y que el valor de σ_Y^2 se conoce con una buena aproximación, se verifica aproximadamente que:

$$\left[\overline{y} - z_{1-\alpha/2}\frac{\sigma_Y}{\sqrt{n}}, \overline{y} + z_{1-\alpha/2}\frac{\sigma_Y}{\sqrt{n}}\right]$$

es un intervalo de confianza para $\overline{Y}$ con coeficiente aproximado $1-\alpha$.

Si se desconoce σ_Y^2, hay que estimarlo por el correspondiente estimador insesgado s^2 y $z_{1-\alpha/2}$ suele reemplazarse por el valor $t_{1-\alpha/2}$ de la distribución t de Student con $n-1$ grados de libertad. Así,

$$\left[\overline{y} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}, \overline{y} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

es un intervalo de confianza con coeficiente $1-\alpha$.

2.8.2. Estimación por intervalo del total

A partir de lo visto para la media, podríamos considerar las aproximaciones anteriores, sin más que multiplicar los extremos por N.

2.8.3. Estimación por intervalo de una proporción

En este caso podría seguirse un razonamiento similar al del muestreo aleatorio sin reposición para construir un procedimiento exacto para la estimación por intervalo de P, sin más que reemplazar la búsqueda en las tablas de la distribución hipergeométrica del caso sin reposición a la búsqueda en las de la binomial $\mathcal{B}(n,P)$, ya que en este caso, ésta última es la que sigue la variable r = número de unidades de la muestra con la propiedad considerada = np. Los cálculos se simplifican bastante con respecto al del muestreo sin reposición, aunque aún resultan algo engorrosos.

También es posible desarrollar un procedimiento aproximado basado en la desigualdad de Tchebychev, de forma que como:

$$\Pr\left(|p-P| \leq \sqrt{\frac{P(1-P)}{n\alpha}}\right) \geq 1-\alpha$$

y

$$|p-P| \leq \sqrt{\frac{P(1-P)}{n\alpha}} \Rightarrow (p-P)^2 \leq \frac{P(1-P)}{n\alpha}$$

$$\Leftrightarrow (n\alpha+1)P^2 - (2pn\alpha+1)P + n\alpha p^2 \leq 0 \Leftrightarrow p_1 \leq P \leq p_2$$

con lo que se verificará que $[p_1, p_2]$ determina un intervalo de confianza para P con coeficiente $1-\alpha$.

Ocasionalmente y cuando no requiere mucha precisión, puede aparecer como intervalo el que se obtiene al reemplazar en la desigualdad de Tchebychev la varianza de P, $\frac{P(1-P)}{n}$ por su estimación insesgada $\frac{p(1-p)}{n-1}$, y nos quedaría el intervalo

$$\left[p-\sqrt{\frac{p(1-p)}{(n-1)\alpha}}, p+\sqrt{\frac{p(1-p)}{(n-1)\alpha}}\right].$$

También puede recurrirse, cuando el tamaño muestral lo permita, a la aproximación normal basada en el Teorema de Moivre según el cual, si n es suficientemente grande (tomaremos $n \geq 30$), se considera válida la aproximación:

$$p \sim \mathcal{N}\left(P, \sqrt{\frac{P(1-P)}{n}}\right)$$

de forma que

$$\Pr\left(|p-P| \leq z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}\right) \simeq 1-\alpha$$

Como

$$|p-P| \leq z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} \Leftrightarrow (p-P)^2 \leq z^2_{1-\alpha/2}\frac{P(1-P)}{n}$$

$$\Leftrightarrow \left(1+\frac{z^2_{1-\alpha/2}}{n}\right)P^2 - \left(2p+\frac{z^2_{1-\alpha/2}}{n}\right)P + p^2 \leq 0 \Leftrightarrow p_1 \leq P \leq p_2$$

El intervalo $[p_1, p_2]$ es un intervalo de confianza para P con coeficiente aproximadamente $1-\alpha$, supuesto el tamaño muestral n suficientemente grande.

También si n es suficientemente grande (tomaremos $n \geq 60$), puede reemplazarse $\frac{P(1-P)}{n}$ por su estimación insesgada $\frac{p(1-p)}{n-1}$, de modo que

$$\left[p-t_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n-1}}, p+t_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n-1}}\right]$$

es otro intervalo de confianza para P con coeficiente aproximadamente $1-\alpha$.

2.9. Elección de tamaños de muestra en el m.a.c.r.

Siguiendo las ideas desarrolladas para el muestreo aleatorio simple, vamos a analizar las principales aproximaciones para determinar los tamaños de muestra adecuados para estimar las características media, total y proporción poblacionales en el muestreo aleatorio con reposición, supuesto que se han especificado una tolerancia e_α y una confianza $1-\alpha$.

MEDIA POBLACIONAL

En principio nuestro objetivo será hallar el menor n tal que

$$\Pr(|\overline{y}-\overline{Y}| \leq e_\alpha) \geq 1-\alpha.$$

Aplicando el Teorema Central del Límite en el supuesto de que lo permite el tamaño muestral (cuando quede $n \geq 30$), entonces puede considerarse que $\overline{y} \sim \mathcal{N}(\overline{Y},\, \sigma_Y/\sqrt{n})$, con lo que reduciríamos el problema a una aproximación de nuestro objetivo que consistiría en determinar el menor n tal que

$$z_{1-\alpha/2}\frac{\sigma_Y}{\sqrt{n}} \leq e_\alpha \Leftrightarrow n \geq \frac{z^2_{1-\alpha/2}\sigma^2_Y}{e^2_\alpha}$$

con lo que, en este caso, n coincide con la aproximación inicial n_0 manejada en el m.a.s. (si $N \to \infty$, $\sigma^2_Y \approx S^2_Y$).

El valor de σ^2_Y suele ser desconocido en la práctica, por lo que puede recurrirse a aproximar su valor por una de las vías señaladas en el muestreo aleatorio simple (estudio piloto, extracción de la muestra en dos pasos, estudios previos sobre características y poblaciones análogas, etc.).

Si se considera la aproximación basada en la desigualdad de Tchebychev, sabemos que

$$\Pr\left(|\overline{y} - \overline{Y}| \leq \sqrt{\frac{\sigma^2_Y}{n\alpha}}\right) \geq 1 - \alpha$$

entonces la aproximación del objetivo que buscamos va a ser determinar el mínimo n tal que

$$\sqrt{\frac{\sigma^2_Y}{n\alpha}} \leq e_\alpha \Leftrightarrow n \geq \frac{\sigma^2_Y}{\alpha e^2_\alpha}$$

con lo que, de nuevo, n coincide con la aproximación inicial n_0 manejada en el m.a.s.

TOTAL POBLACIONAL

Es evidente a partir del estudio para la media.

PROPORCIÓN POBLACIONAL

En la estimación de la proporción pueden seguirse en buena parte las ideas expuestas para el muestreo aleatorio simple.

Si el tamaño muestral lo permite, es decir, se supone que resultará suficientemente grande como para aplicar el Teorema de Moivre, tendríamos que:

$$\Pr\left(|p - P| \leq z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}\right) \simeq 1 - \alpha$$

de modo que

$$z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} \leq e_\alpha \Rightarrow n \geq \frac{z^2_{1-\alpha/2}P(1-P)}{e^2_\alpha}$$

con lo que, n coincide de nuevo con la aproximación inicial n_0 manejada en el m.a.s.

En la expresión última figura el producto $P(1-P)$ que se supone desconocido y puede aproximarse por alguno de los caminos ya sugeridos en el muestreo aleatorio simple (el conservador que aproxima $P(1-P)$ por su mayor valor posible de acuerdo con la información disponible o el basado en un muestreo en dos etapas).

En cuanto al procedimiento basado en la aproximación por la desigualdad de Tchebychev, teniendo en cuenta que de acuerdo con ella se cumple

$$\Pr\left(|p-P| \leq \sqrt{\frac{P(1-P)}{n\alpha}}\right) \geq 1-\alpha$$

si tomamos un n tal que

$$\frac{P(1-P)}{n\alpha} \leq e_\alpha^2 \Rightarrow n \geq \frac{P(1-P)}{\alpha e_\alpha^2}$$

con lo que, de nuevo, n coincide con la aproximación inicial n_0 manejada en el m.a.s.

2.10. Cálculo del tamaño muestral en el m.a.s. y m.a.c.r.

La fórmula general para el cálculo del tamaño de muestra necesario para cometer un error máximo admisible o tolerancia, d, es

- en poblaciones infinitas (m.a.c.r.)

$$n_0 = \frac{K^2}{d^2}$$

- en poblaciones finitas de tamaño N (m.a.s.),

$$\underline{\text{Media y total poblacional}} \qquad n = \frac{n_0}{1+\frac{n_0}{N}} \qquad \underline{\text{Proporción poblacional}} \qquad n = \frac{n_0}{1+\frac{(n_0-1)}{N}}$$

El valor de K^2 varía según el parámetro poblacional que se quiere estimar

- media poblacional, $\bar{Y}$,
- total poblacional, Y_T, (sólo en poblaciones finitas)
- proporción poblacional, P,

y según la condición impuesta, que puede ser:

- error de muestreo fijado, $d = e$,
- error relativo de muestreo fijado, $d = e_r$,
- error de muestreo y coeficiente de confianza fijados, $d = e_\alpha$,
- error relativo de muestreo y coeficiente de confianza fijados, $d = e_{r\alpha}$.

	Valor de K^2	$\bar{Y}$	Y_T	P
e		S_Y^2	$N^2 S_Y^2$	$P(1-P)$
e_r		C_Y^2	C_Y^2	$\dfrac{1-P}{P}$
e_α	si Aprox. normal	$z_{1-\alpha/2}^2 S_Y^2$	$z_{1-\alpha/2}^2 N^2 S_Y^2$	$z_{1-\alpha/2}^2 P(1-P)$
e_α	si Tchebychev	$\dfrac{S_Y^2}{\alpha}$	$\dfrac{N^2 S_Y^2}{\alpha}$	$\dfrac{P(1-P)}{\alpha}$
$e_{r\alpha}$	si Aprox. normal	$z_{1-\alpha/2}^2 C_Y^2$	$z_{1-\alpha/2}^2 C_Y^2$	$z_{1-\alpha/2}^2 \dfrac{1-P}{P}$
$e_{r\alpha}$	si Tchebychev	$\dfrac{C_Y^2}{\alpha}$	$\dfrac{C_Y^2}{\alpha}$	$\dfrac{(1-P)}{P\alpha}$

donde

S_Y^2 = cuasivarianza poblacional ($\approx \sigma_Y^2$ en poblaciones infinitas)

C_Y = coeficiente de variación = $\dfrac{S_Y}{\bar{Y}}$.