Estadísticas para las ciencias contables: Extracción de la muestra en los diferentes diseños muestrales y cálculo de su tamaño*

Tulio Echeverri Uribe

Economista Universidad de Antioquia, Magíster Diplome D`études Aprofondies (D.E.A.) - Universidad Sorbona, París
Doctorat de Troisieme Cycle - Universidad Sorbona, París
Coautor del Manual para el control selectivo en la Contraloría del Departamento de Antioquia
Coautor del Estudio sobre la satisfacción del cliente en las Empresas Públicas de Medellín
Profesor tiempo completo, Universidad de Antioquia
tgecheverri@agustianianos.udca.edu.co

* Este artículo hace parte del libro del mismo autor titulado "Estadísticas para las ciencias contables" en proceso de publicación.

ESTADÍSTICAS PARA LAS CIENCIAS CONTABLES: EXTRACCIÓN DE LA MUESTRA EN LOS DIFERENTES DISEÑOS MUESTRALES Y CÁLCULO DE SU TAMAÑO

Resumen: Además del uso generalizado que en la auditoría se hace del muestreo estadístico, en Colombia la Constitución Nacional en su artículo 267 establece que el control fiscal será ejercido en forma posterior y selectiva. Para cumplir esta norma constitucional, el auditor debe estar capacitado para responder las tres grandes preguntas del muestreo selectivo: ¿Cuál debe ser el tamaño de la muestra aleatoria? ¿Cómo se extrae la muestra selectiva? ¿Qué conclusiones se sacan de la muestra extraída?. Este artículo da elementos estadísticos a los auditores para responder las dos primeras preguntas y así garantizar que la muestra extraída sea representativa de la población y asegurar resultados efectivos y confiables en el control fiscal, así como en los demás controles que se apoyen en el muestreo estadístico.

Palabras clave: inferencia estadística, diseño muestral, clases de muestreo probabilístico, muestreo intencional, muestreo sin norma, tamaño de la muestra, auditoría, control interno, control fiscal.

STATISTICS FOR ACCOUNTING SCIENCES: EXTRACTION OF THE SAMPLE IN THE DIFFERENT SAMPLES DESIGNS AND THE CALCULATION FROM ITS SIZE

Summary: In addition to generalized use performed in auditing to statistical sampling, the Colombian National Constitution, Article 267, establishes that the fiscal control will be exerted in a selective and former manner. In order to fulfill this constitutional norm, the auditor must be enabled to respond to the three great questions of the selective sampling: Which must be the random sample size? How is the selective sample extracted? What conclusions can be drawn from the extracted sample?. This article gives statistical elements to the auditors to respond to the first two questions and thus guarantee that the extracted sample be representative of the population and to assure effective and reliable results in the fiscal control, as well as in the other controls that lean on the statistical sampling.

Key words: statistical inference, sample design, classes of probabilistic sampling, intentional sampling, sampling without norm, sample size, audit, internal control, fiscal control.

STATISTIQUES POUR LES SCIENCES COMPTABLES: EXTRACTION DE L ' ÉCHANTILLON DANS LES DIFFÉRENTES CONCEPTIONS D ' ÉCHANTILLONNAGE ET LE CALCUL DE SA TAILLE

Résumé: En plus de l'utilisation généralisée qui dans l'audit est faite de l'échantillonnage statistique, la Constitution Nationale Colombienne établit dans son article 267 que le contrôle fiscal sera exercé de manière postérieure et sélective. Pour respecter cette norme constitutionnelle, le réviseur des comptes doit être formé pour répondre les trois grandes questions de l'échantillonnage sélectif : Quel doit être la taille de l'échantillon aléatoire? Comment extrait-il l ' échantillon sélectif? Quelles conclusions sont tirées de l'échantillon extrait? Cet article donne des éléments statistiques aux réviseurs des comptes pour répondre les deux premières questions et ainsi garantir que l'échantillon extrait soit représentatif de la population et d'assurer des résultats effectifs et fiables dans le contrôle fiscal, ainsi que les autres contrôles qui s'appuient dans l'échantillonnage statistique.

Mots clef: inférence statistique, plan d ' échantillonnage, classes d'échantillonnage probabiliste, échantillonnage intentionnel, échantillonnage sans norme, taille de l'échantillon, audit, contrôle interne, contrôle fiscal.

Estadísticas para las ciencias contables: Extracción de la muestra en los diferentes diseños muestrales y cálculo de su tamaño

Tulio Echeverri Uribe

Primera versión recibida: agosto de 2004; versión final aceptada: marzo de 2005 (Eds.)

Introducción

A la pregunta si el profesional de las Ciencias Contables, debe o no tener conocimientos en Estadística, la respuesta a dar es positiva y aún más, se debe enfatizar en que debe tener conocimientos excelentes en esta disciplina. La justificación de la anterior respuesta se encuentra en las siguientes razones:

- En la economía moderna, al contador se le exige tener un perfil gerencial para lo cual es necesario que tenga buenos conocimientos en Estadística.
- Una notoria opción profesional del contador, es la auditoría, la cual requiere excelentes conocimientos en muestreo y en inferencia estadística.
- La Constitución Nacional en su artículo 267 establece que el control fiscal será ejercido en forma posterior y selectiva.

Ahora, en el control selectivo se hacen tres grandes preguntas: *Primera:* Cómo se extrae la muestra aleatoria. *Segunda:* Cuál debe ser el tamaño de la muestra aleatoria. *Tercera:* Para qué sirve el resultado muestral. En este artículo nos ocuparemos de dar respuesta a las dos primeras preguntas y en un artículo posterior se dará respuesta a la tercera pregunta. Para responder a la pregunta: *cómo se extrae la muestra aleatoria,* después de hacer algunas consideraciones sobre la técnica del muestreo, consideraremos la manera de extraer la muestra en los diferentes diseños muestrales probabilísticos.

I. Algunas consideraciones sobre la técnica del muestreo

1. El concepto de inferencia estadística y algunas definiciones de diferentes términos técnicos utilizados en el muestreo

Un objetivo muy importante de la estadística es hacer inferencia acerca de una población con base en la información contenida en una muestra. En consecuencia, definamos en primer lugar, el concepto de inferencia estadística.

Definición 1: La inferencia estadística es un proceso de obtener información para toda la población a partir de la muestra añadiendo el grado de confianza que se tiene en esa información. En efecto, las características de una población, que se conocen con el nombre de parámetros, generalmente no son conocidas directamente; en cambio serán estimadas de acuerdo a los resultados del muestreo. Así, un auditor por ejemplo, extrae una muestra de n registros para estimar el porcentaje de facturas mal codificadas en una empresa. Del mismo modo un contador toma una muestra de n facturas para estimar el promedio y el total de las ventas de la compañía.

Definición 2: Un elemento o unidad de muestreo es un objeto en el cual se toman mediciones. Así, un elemento sería una factura y su medición podría ser de la siguiente manera: *cero* si la factura está mal codificada y *uno* si está bien codificada.

Definición 3: Una población es una colección de elementos acerca de los cuales deseamos hacer alguna inferencia. Una población sería, por ejemplo, el conjunto de facturas de la compañía y la característica de interés en cada elemento de la población es su codificación. Una tarea importante para el auditor es definir cuidadosamente y completamente la población antes de extraer la muestra y su definición debe contener una descripción de los elementos que serán incluidos y una especificación de las mediciones que se van a considerar.

Definición 4: Un marco muestral es una lista de elementos o unidades de muestreo.

Definición 5: Una muestra es una colección de unidades seleccionadas de un marco o de varios marcos muestrales. Es de anotar que la muestra debe ser representativa de la población lo mejor posible para garantizar la precisión y la confiabilidad de la inferencia.

Definición 6: Muestreo es una técnica mediante la cual se obtiene la muestra.

2. Clases de Muestreo

El muestreo puede ser: Probabilístico, a juicio o intencional, sin norma. Es **Probabilístico:** cuando cada elemento de la población tiene la misma probabilidad de formar parte de la muestra aleatoria. En la elección de la muestra en este muestreo no interviene para nada el auditor.

A juicio o intencional: Cuando hay subjetividad en la extracción de la muestra. El auditor determina previamente los criterios que debe cumplir un elemento de la población para hacer parte de la muestra. Así, si el auditor sospecha, por ejemplo, que cierto funcionario ha hecho fraude en la facturación de una entidad, incluirá en la muestra los documentos procesados por este funcionario.

Sin norma: Cuando la muestra se toma sin ningún criterio.

Observación: solo el muestreo probabilístico se presta para hacer inferencia estadística.

3. Necesidad del muestreo

En las siguientes situaciones es mejor estudiar la población a través de muestras en lugar de hacer un censo, es decir, en lugar de medir todas las unidades poblacionales:

- Cuando la población es infinita o tan grande que el costo y el tiempo hacen imposible el censo.
- Cuando existe una gran homogeneidad en las unidades poblacionales que hace que una muestra aleatoria contenga la información de interés necesaria en el estudio.
- Cuando el proceso de medida o de investigación de la característica en estudio es de tipo destructivo lo cual necesariamente obliga al conocimiento de la población a través de una muestra. Por ejemplo el examen de fósforos en un control de calidad.

4. Ventajas del muestreo frente al censo

- La conveniencia fundamental del muestreo se encuentra en los bajos costos y en la economía del tiempo y del personal humano. En general, la calidad de la información obtenida a través de una muestra selectiva es superior a la obtenida en un censo. En efecto, cuando se realiza un muestreo el control que puede ejercerse sobre la recolección y procesamiento de la información es muy superior al que se tiene cuando se realiza un censo.
- Los errores, especialmente los no muestrales disminuyen considerablemente cuando se realiza una muestra selectiva. Errores no muestrales son por ejemplo: Las no respuestas, las inconsistencias, las codificaciones erradas.

5. Problemas prácticos que se presentan en el muestreo

Los siguientes son los principales problemas del muestreo:

- Encuesta mal elaborada debido a las siguientes causas:

a. La encuesta no mide exactamente lo que se quiere conocer, es decir, el objetivo inicial no es captado con claridad.

b. La encuesta contiene preguntas mal elaboradas las cuales no cubren las alternativas posibles y además presentan redacción confusa.

c. En la encuesta faltan preguntas esenciales.

d. La encuesta es demasiado compleja y extensa y además los encuestadores no la dominan.

- Problemas en el procesamiento de la información. Se puede presentar en las siguientes circunstancias:

a. Mala codificación de las preguntas, lo cual dificulta el empleo del computador.

b. No hubo acuerdo previo con el programador y en consecuencia algunas preguntas requieren de procesos complejos para el procesamiento de la información.

c. Mala elaboración de los cruces entre preguntas y los cuadros de salida.

- Problemas en los encuestadores. Existen problemas en los encuestadores cuando:

a. No dominan suficientemente la encuesta.

b. Después de conocer aproximadamente las repuestas, los encuestadores llenan las encuestas arbitrariamente sin visitar y sin investigar la unidad de muestreo.

c. Para ganar más dinero hacen parte de la encuesta y la otra la llenan arbitrariamente.

d. No han sido suficientemente entrenados.

- Problemas administrativos cuando se presenta:

a. Carencia de un plan o cronograma para ir desarrollando todas las etapas del muestreo.

b. Carencia o mal diseño de planillas para ir controlando la labor de los encuestadores.

- Problemas en el tamaño de la muestra. En el muestreo aparecen problemas cuando para determinar el tamaño de la muestra se procede arbitrariamente.

6. Soluciones a los problemas que se presentan en el muestreo

Las principales soluciones son:

- Obtención de una muestra piloto, la cual es la realización completa del muestreo en pequeña escala, es decir, se hace una prueba total de 25 a 50 unidades, donde se realiza la encuesta, se procesan los datos y se analizan como si se tratara del trabajo final.

La muestra piloto tiene las siguientes ventajas:

* Permite corregir a tiempo todas las dificultades del diseño muestral.
* Permite superar todas las dificultades del procesamiento de la información.
* Familiariza a los investigadores con la magnitud de las cifras, lo que permite en el muestreo real, controlar las respuestas.
* Permite calcular el tamaño de la muestra al proporcionar una estimación de los parámetros poblacionales necesarios para hallar el tamaño muestral.
* Permite diseñar los cuadros de salida para ir trasladando la información a medida que se produce.

Es de anotar, que la realización de un muestreo piloto es sin duda la operación más importante del muestreo.

- Proporcionar una preparación estricta y adecuada a los encuestadores.

- Mantener una supervisión de los encuestadores.

No hay que olvidar que la materia prima del muestreo la manejan personas que en general tienen una vinculación muy incidental con el mismo. Por tanto debe haber una supervisión eficiente sobre los encuestadores. La regla general es: toda encuesta debe pasar por una etapa de control.

La supervisión comprende:

a. Una primera revisión en la recepción de la encuesta ya elaborada.
b. Una comprobación personal, directamente en el sitio en que el encuestador debió realizar la encuesta. De esta forma se comprueba si el encuestador estuvo allí, se verifica el tiempo de permanencia y se pueden repetir dos ó tres preguntas para contrastarlas con las respuestas consignadas en la encuesta. Lo ideal es controlar todas las encuestas, pero si esto no es posible, se deben revisar las que permitan las circunstancias.

- Hacer la elaboración de la encuesta con la participación del experto en computadores.
- Mantener buenos controles administrativos.

II. Clases de muestreo probabilístico y forma de extracción de la muestra en cada diseño muestral

El procedimiento para seleccionar la muestra aleatoria se denomina diseño de muestreo o procedimiento de muestreo o técnicas de muestreo. Para garantizar la calidad de la información que proporcione estimadores precisos de los parámetros poblacionales es necesario que el auditor elija el diseño de muestreo adecuado para obtener la muestra. Analizaremos los siguientes diseños fundamentales de muestreo:

- Muestreo con reemplazamiento o muestreo de población infinita.
- Muestreo sin reemplazamiento o muestreo aleatorio simple o muestreo de población finita.
- Muestreo aleatorio sistemático.
- Muestreo sistemático en poblaciones periódicas.
- Muestreo sistemático «especial».
- Muestreo sistemático replicado.
- Muestreo aleatorio estratificado.
- Muestreo aleatorio por conglomerados.
- Muestreo aleatorio por conglomerados en dos etapas.
- Muestreo con inclusión forzosa.

Veamos el estudio detallado de cada una de éstas técnicas de muestreo y la manera de extraer la muestra aleatoria en cada diseño.

1. **Muestreo con reemplazamiento (M.C.R.) ó muestreo de población infinita.** Se obtiene reponiendo el elemento a la población después de cada observación. En este diseño muestral todas las unidades de la población tienen la misma probabilidad de ser seleccionadas y las observaciones son independientes.

2. **Muestreo sin reemplazamiento ó muestreo aleatorio simple (M.A.S.) ó muestreo de población finita.** Se obtiene cuando el elemento poblacional observado no es devuelto a la población. En esta técnica podemos dar la siguiente definición:

Definición 7: Si un tamaño de muestra n es seleccionada de una población de tamaño N de tal manera que cada muestra posible de tamaño n tiene la misma probabilidad de ser seleccionada, el procedimiento de muestreo se denomina muestreo aleatorio simple.

En esta técnica de muestreo las observaciones son dependientes. Para el análisis estadístico se tiene en cuenta lo siguiente:

- Se considera muestreo de población infinita o M.C.R. cuando las extracciones de las unidades de la población se hacen con sustitución o cuando las extracciones se hacen sin sustitución pero el tamaño de la muestra con relación al tamaño poblacional es muy pequeño, considerando una muestra pequeña cuando $n \leq 0.05N$.
- Se considera muestreo de población finita o M.A.S. cuando las extracciones de las unidades poblacionales se hacen sin sustitución pero el tamaño de la muestra es grande con relación al tamaño de la población y se considera una muestra grande cuando $n > 0.05N$.

En resumen, en la práctica las extracciones de los elementos poblacionales se hace sin sustitución como es lógico; por tanto la diferencia entre el muestreo de población infinita y el muestreo de población finita radica en la relación del tamaño de la muestra con respecto al tamaño de la población.

Tanto el muestreo de población infinita como el de población finita suministran una buena estimación de los parámetros poblacionales cuando la población es homogénea. Además, estas dos técnicas forman la base de la mayoría de los diseños muestrales.

- Cómo seleccionar la muestra aleatoria en el muestreo de población infinita y en el muestreo de población finita

La condición esencial del muestreo es que la selección de los elementos de la muestra debe ser aleatoria. Esta es una condición indispensable para hacer uso correcto de los procesos de estimación estadísticos. Por tanto es necesario el uso de un mecanismo de probabilidad bien diseñado para la selección de la muestra. El uso de éste mecanismo es el que distingue las muestras probabilísticas de las pruebas subjetivas o a juicio. Uno de los procedimientos más utilizados es seleccionar la muestra a través de la generación de números aleatorios. Para ello, después de tener una correcta identificación de todos los elementos de la población, ésta se enumera y se generan n números aleatorios y los correspondientes elementos muestrales constituyen la muestra seleccionada. En el muestreo de población finita, si un número generado aparece de nuevo, se elimina y se genera otro. Un error del auditor sería descartar una muestra seleccionada aleatoriamente porque no le «parece correcta», ya que así destruye el carácter probabilístico de la muestra y el auditor terminaría seleccionando una muestra subjetiva.

Los números aleatorios los podemos generar en cualquier paquete estadístico. En este artículo, para las aplicaciones informáticas, utilizaremos el paquete estadístico Statgraphics Plus, cuyo manejo no ofrece ninguna dificultad.

Para generar los números aleatorios los comandos son:

Comandos

Se elige la opción DESCRIBE, luego DISTRIBUTIONS y PROBABILITY DISTRIBUTIONS; aquí se elige la distribución con la cual se desea generar los números aleatorios. Si se quiere generar números enteros, elegimos la distribución DISCRETE UNIFORM. Luego en el icono amarillo de las Tabular Options seleccionamos RANDOM NUMBERS y con clic derecho en Pane Options en SIZE escribimos la cantidad de números aleatorios que se desea generar y en Analysis Options en LOWER LIMIT y en UPPER LIMIT se registra el rango en el cual deben estar los números generados. Estos deben ser salvados de la siguiente manera: en el cuarto icono en Save Result Options activar SAVE y nombrarles como se desee. Si se quiere mirarlos se debe abrir el archivo por WINDOW y elegir la opción número 1.

Ejemplo. Se desea analizar algunas características de los empleados de la sección de salud departamental seleccionando una muestra de 20 empleados. Para identificar la población lo hacemos a través de un marco muestral conveniente que sería la lista de los 200 empleados de la sección de salud, ya que un empleado es una unidad muestral individual. Para extraer la muestra generamos 20 números aleatorios sin repetición del 1 al 200 y así obtenemos los 20 empleados que conforman la muestra.

3. **Muestreo aleatorio sistemático.** Es un diseño de muestreo que es ampliamente utilizado debido principalmente a que simplifica el proceso de selección de la muestra y se constituye en una alternativa del muestreo aleatorio simple ya que es más fácil de llevar a cabo y por tanto está menos expuesto a los errores del auditor. Además frecuentemente proporciona más información que el muestreo aleatorio simple por unidad de costo.

La idea básica del muestreo sistemático es la siguiente: calcular el valor de la constante k definida como $k = \frac{N}{n}$, donde N: tamaño de la población y n: tamaño de la muestra. Luego, se selecciona aleatoriamente un elemento de los primeros k elementos poblacionales, el cual sería el punto de inicio aleatorio y posteriormente se hace la selección de cada k–ésimo elemento de la población hasta completar el tamaño de la muestra requerida. Es de anotar que para aplicar el muestreo sistemático no es indispensable que la población esté enumerada.

Esta técnica la podemos definir de la siguiente manera:

Definición 9: Una muestra obtenida al seleccionar aleatoriamente un elemento de los primeros k elementos poblacionales y luego la selección de cada k–ésimo elemento se denomina una muestra sistemática de uno en k, donde el valor de k se obtiene de la relación $k = \frac{N}{n}$.

En algunos casos es preferible utilizar el muestreo sistemático y no el muestreo aleatorio simple como lo podemos ver en la siguiente situación:

Supongamos que la mayoría de los primeros 2000 recibos de ventas de una compañía han sido archivados correctamente pero a causa de un cambio de empleado los siguientes 2000 recibos han sido archivados incorrectamente en su mayoría. El muestreo aleatorio simple podría accidentalmente seleccionar un gran número de recibos de los primeros 2000 y seleccionar muy pocos del segundo grupo de recibos y en consecuencia se produciría una estimación muy deficiente del porcentaje de recibos mal archivados. En contraste, el muestreo aleatorio sistemático seleccionaría un número igual de recibos de cada uno de los dos grupos y así se tendría una estimación más precisa del porcentaje de recibos mal archivados.

Para decidir si se aplica el muestreo aleatorio simple o el muestreo aleatorio sistemático es necesario considerar el tipo de población que se investiga. Consideremos tres tipos de poblaciones: Población aleatoria, población ordenada, población periódica. Analicemos estos tres tipos de poblaciones.

- Población aleatoria. Su definición es la siguiente:

  Definición 10: Una población es aleatoria sí sus elementos están ordenados al azar. Cuando la población tiene ordenación aleatoria, una muestra sistemática es equivalente a una muestra aleatoria simple ya que las varianzas del estimador del valor medio poblacional son prácticamente iguales en los dos diseños muestrales. En consecuencia, en una población aleatoria es indiferente aplicar el muestreo aleatorio simple o el muestreo sistemático para la extracción de la muestra.

  Población ordenada.

  Definición 11: Una población es ordenada si los elementos dentro de la población están ordenados en magnitud de acuerdo con algún esquema.

Un ejemplo sería el siguiente: ordenar de acuerdo a su valor ascendente las facturas de servicios públicos del estrato III.

En una muestra sistemática extraída de una población ordenada, generalmente la varianza del estimador del promedio es menor que la varianza del estimador de una muestra aleatoria simple. En consecuencia, en las poblaciones ordenadas es preferible aplicar el diseño muestral sistemático y no el aleatorio simple.

Población periódica

Definición 12: Una población es periódica si sus elementos tienen variación cíclica.

Un ejemplo sería el siguiente: Por muestra selectiva se quiere estimar el promedio de las ventas diarias de un supermercado. Claramente la población de las ventas diarias del supermercado es periódica, ya que las ventas mayores se presentan al final de la semana. Si se aplica el muestreo sistemático y se muestrea por ejemplo cada miércoles, probablemente se subestimaría el verdadero promedio de las ventas y si se muestrea cada sábado probablemente se sobrestimaría el promedio verdadero de las ventas diarias.

En consecuencia, en las poblaciones periódicas es mejor aplicar el muestreo aleatorio simple que el muestreo sistemático ya que la varianza del estimador del promedio poblacional es menor en el diseño aleatorio simple que en el diseño sistemático.

4. Muestreo sistemático en poblaciones periódicas. En las poblaciones periódicas si se quiere aplicar el muestreo sistemático es necesario cambiar varias veces el punto de inicio de la muestra aleatoria, lo cual reduce la posibilidad de seleccionar observaciones con la misma posición relativa y así se corregiría la variación perturbadora cíclica de la información. Por ejemplo, en una población periódica para romper las variaciones cíclicas, el auditor decide en un muestreo sistemático cambiar el punto de inicio después de cada 15 elementos extraídos; así si por ejemplo, $k = 10$ y el punto de inicio aleatorio es 5, los 15 primeros elementos seleccionados son: $5, 15, 25, \cdots, 155$ en este momento el auditor debe seleccionar otro punto de inicio aleatorio entre 155 y 165; si es seleccionado aleatoriamente el 161 por ejemplo, los siguientes 15 elementos serían: $161, 171, 181 \cdots$ este procedimiento se repite hasta obtener el tamaño de muestra requerido. Si se aplica este procedimiento de seleccionar varias veces el punto de inicio aleatorio en una población periódica, podemos suponer que la muestra obtenida es equivalente a una muestra sistemática extraída de una población aleatoria, ya que en gran parte se ha eliminado la variación cíclica.

Es de anotar que éste procedimiento no es estrictamente aleatorio ya que la decisión de cambiar de punto de inicio después de la extracción de 15 elementos poblacionales, por ejemplo, es una decisión a juicio del auditor.

Cómo seleccionar una muestra sistemática aleatoria

El procedimiento es el siguiente: se debe calcular el valor de k, donde la constante k se obtiene así: $k = \frac{N}{n}$ Ahora, para definir el punto de inicio de extracción de la muestra, se genera un número aleatorio entre 1 y k y luego cada k–ésimo elemento es seleccionado uno hasta completar los n elementos que constituyen la muestra.

Así, por ejemplo, si se tienen 5000 facturas y el tamaño de la muestra es 100 facturas, se procede así. El valor de k es:

$$k = \frac{5000}{100} = 50$$

Luego, se genera un número aleatorio entre 1 y 50; si por ejemplo se generó el 15, las facturas elegidas son las que poseen las siguientes posiciones: La 15, la 65, la 115, ... hasta extraer las 100 facturas que constituyen la muestra.

Muestreo sistemático replicado

Este diseño muestral consiste en hacer varias réplicas de una muestra aleatoria sistemática básica, o sea, se seleccionan varias muestras sistemáticas. En la práctica, se ha comprobado que el número de muestras sistemáticas que se deben seleccionar, deben ser alrededor de 10, es decir, $n_s = 10$, las cuales van a permitir obtener suficientes medias muestrales para obtener una estimación satisfactoria de los parámetros poblacionales. Este diseño muestral no necesita hacer ningún supuesto sobre la naturaleza de la población, lo cual constituye su gran ventaja y lo convierte en una excelente técnica para obviar la perturbadora variación cíclica de las poblaciones periódicas. Además, en este diseño muestral se logra obtener una muestra con representación adecuada de la población, lo cual es la condición esencial del éxito en el muestreo. Para aplicar el diseño muestral sistemático replicado, se dan los siguientes pasos:

1. Calcular el valor de k, así $k = \dfrac{N}{n}$,
2. Calcular el valor de k' donde $k' = 10 * k$
3. Generar 10 números aleatorios entre 1 y k', los cuales serán los puntos de inicio aleatorio para las 10 muestras sistemáticas.
4. El segundo elemento de cada muestra sistemática se obtiene adicionando k' al primer elemento; el tercer elemento se obtiene adicionando k' al segundo elemento y así sucesivamente hasta completar el tamaño de la muestra.

Apliquemos esta técnica de muestreo en el siguiente ejemplo.

Ejemplo. Un auditor quiere hacer un control selectivo en las 420 facturas del servicio de Internet de una compañía, y para obtener la muestra de 70 facturas aplica el muestreo sistemático replicado.

Los pasos a seguir son:

1.Calcular el valor de k: $k = \frac{N}{n} = \frac{420}{70} = 6$

2.Calcular el valor de k': $k' = 10 * k = 10 * 6 = 60$

3.se generan 10 números aleatorios entre 1 y 60 y, por ejemplo, los resultados fueron: 11, 38, 3, 42, 23, 8, 56, 47, 50, 32 estos números son los puntos de inicio aleatorios de las 10 muestras sistemáticas de 7 elementos cada una.

En la siguiente tabla se muestran los números aleatorios para formar las 10 muestras sistemáticas.

	Punto de inicio aleatorio	Segundo	Tercero	Cuarto	Quinto	Sexto	Séptimo
Muestra 1	3	63	123	183	243	303	363
Muestra 2	8	68	128	188	248	308	368
Muestra 3	11	71	131	191	251	311	371
Muestra 4	23	83	143	203	263	323	383
Muestra 5	32	92	152	212	272	332	392
Muestra 6	38	98	158	218	278	338	398
Muestra 7	42	102	162	222	282	342	402
Muestra 8	47	107	167	227	287	347	407
Muestra 9	50	110	170	230	290	350	410
Muestra 10	56	116	176	236	296	356	416

Si se analizan los resultados de las 10 muestras sistemáticas se puede concluir que la repartición de los números de las facturas que forman la muestra es excelente lo cual garantiza que la muestra sea representativa de la población.

6. **Muestreo sistemático especial**: En algunos procesos aleatorios no es posible conocer el tamaño poblacional, N, o es imposible calcularlo ya que la población está siempre en continua formación. Algunos ejemplos serían: El control, por muestra selectiva, de los automotores que pasan por un peaje en un día festivo. El extraer una muestra aleatoria, para el control de calidad en la línea de producción de una compañía cervecera. El seleccionar una muestra aleatoria para analizar la satisfacción del cliente en las

personas que visitan un supermercado en un día cualquiera. En ninguno de los tres ejemplos es posible calcular el tamaño poblacional y por tanto no es posible aplicar el muestreo aleatorio simple donde la población debe estar debidamente enumerada y tampoco es posible aplicar el muestreo aleatorio sistemático ya que no se puede conocer el valor de k. En estos procesos se aplica el muestreo sistemático «especial» cuya técnica es la siguiente:

El auditor decide a su juicio el valor de la constante k; luego, genera un número aleatorio entre 1 y k el cual es el punto de inicio aleatorio para la muestra sistemática y posteriormente cada k–ésimo elemento poblacional es seleccionado hasta completar el tamaño de la muestra requerida. Como se puede observar, el valor que el auditor le asigna a k es un elemento subjetivo, no aleatorio en ésta técnica de muestreo. Supongamos, por ejemplo, que un auditor quiere analizar la satisfacción de los usuarios del transporte masivo bajo la modalidad *metro* de alguna ciudad. Para ello decide tomar una muestra de n usuarios que utilizan el metro de 6 a 7 de la mañana, en la estación principal y que salen por la puerta Sur. Para obtener la muestra aplica el diseño de muestreo sistemático «especial» ya que la población está en continua formación. Si el auditor desea una selección de usuarios muy próxima decide un valor bajo de k, por ejemplo 5 y genera un número aleatorio entre 1 y 5; si se generó el 2, los usuarios que salen por la puerta Sur en el orden 2, 7, 12, 17, 22, ... serán los integrantes de la muestra. Si el auditor desea una selección más separada de los usuarios, decide un valor mayor para la constante k, por ejemplo, de 20 y aplica la técnica del muestreo sistemático.

7. **Muestreo aleatorio estratificado.** Este diseño muestral se aplica cuando la población es heterogénea y se puede separar las unidades de la población en subgrupos llamados estratos. En muchos procesos de auditoria es necesario estratificar la población. Por ejemplo, en una auditoria de los ingresos de un municipio es necesario hacer un muestreo selectivo estratificado ya que se tendrían facturas de valores muy pequeños y facturas de valores muy elevados. En algunos casos la estratificación se presenta de manera evidente, como es el caso de los estudios de carácter espacial donde las subdivisiones geográficas se pueden asemejar a estratos. En otras situaciones, la estratificación se basa en una o más variables de interés sobre las cuales se posee suficiente información. Estas variables pueden ser cuantitativas tales como: edad, ingreso, peso,

longitud, etcétera o pueden ser cualitativas tales como: sexo, estado civil, filiación política, etcétera. Con la estratificación se procura que exista una gran homogeneidad dentro de las unidades correspondientes a cada estrato y que exista una gran heterogeneidad entre los diferentes estratos. En este diseño muestral podemos dar la siguiente definición:

Definición 13: Una muestra aleatoria estratificada es la obtenida mediante la separación de los elementos poblacionales en grupos llamados estratos de tal manera que cada elemento pertenezca a uno y solo uno de los estratos y luego se hace la selección posterior de una muestra aleatoria de cada estrato.

Ventajas del muestreo aleatorio estratificado

- Si la estratificación se realiza inteligentemente, los estimadores de los parámetros poblacionales obtenidos en el muestreo aleatorio estratificado son más eficientes que los obtenidos en el muestreo aleatorio de población infinita o de población finita.
- Como la variabilidad de las observaciones dentro de cada estrato es más pequeña que la variabilidad de toda la población, éste método permite hacer estimaciones para subgrupos de la población sin necesidad de seleccionar otra muestra y por tanto sin ningún costo adicional.
- Por la partición de una población grande y heterogénea en estratos, el costo de recolectar y analizar la información es reducido.

Cómo seleccionar una muestra en el diseño muestral aleatorio estratificado

Una muestra aleatoria estratificada se obtiene separando los elementos de la población en grupos o estratos de tal manera que cada elemento pertenezca a uno y solo uno de los estratos y luego se elige una muestra aleatoria con muestreo de población infinita o de población finita o con muestreo sistemático de manera independiente en cada estrato, es decir, la muestra debe ser seleccionada de tal manera que las observaciones elegidas en un estrato no dependen de las que han sido seleccionadas en otro estrato.

Cómo formar los estratos

El éxito del diseño muestral estratificado radica en la formación inteligente de los estratos, de tal manera que cada estrato sea homogéneo en su interior, pero muy diferente con respecto a los otros estratos. En la estratificación hay que resolver dos problemas: ¿Cuál es el número de estratos a formar? ¿Cuál es la dimensión de cada estrato?. Para ello, veamos dos métodos importantes para la formación de los estratos:

Primero: el display de tallo y hojas.
Segundo: el método acumulativo de la raíz cuadrada de la frecuencia absoluta.

Analicemos cada uno de estos métodos. El display de tallo y hojas: «Stem and leaf display». Este método gráfico es una gran ayuda para decidir en una estratificación el número de estratos y la delimitación precisa de cada estrato. Veámoslo en el siguiente ejemplo:

Ejemplo. Un auditor desea estimar el promedio de las ventas diarias de 56 empresas, a través de un muestreo aleatorio estratificado. De acuerdo a los ingresos, las ventas diarias de las 56 empresas se agruparon en intervalos de la siguiente manera:

La tabla de distribución de frecuencias de los ingresos diarios de los 56 almacenes es la siguiente:

Ventas diarias	*Marca de clase*	*Frecuencia absoluta*
1.000.000–1.500.000	1.250.000	11
1.500.000–2.000.000	1.750.000	14
3.000.000–3.500.000	3.250.000	9
3.500.000–4.000.000	3.750.000	4
5.000.000–5.500.000	5.250.000	5
5.500.000–6.000.000	5.750.000	8
6.000.000–6.500.000	6.250.000	3
6.500.000–7.000.000	6.750.000	2

Como el coeficiente de variación de las ventas diarias es del 45.5352%, esta información es heterogénea y debe estar estratificada y para ello vamos a analizar el display de tallo y hojas.

En el Statgraphics Plus para obtener el Display de Tallo y Hojas los comandos son: se parte de la opción DESCRIBE luego Numeric Data y One–Variable Analysis.

Se activa la variable en estudio y en el icono amarillo de las Tabular Options se elige Stem and Leaf Display. La salida del computador del Stem and Leaf Display en este ejemplo es:

```
Stem-and-Leaf Display for Ventas: unit = 100000.0   1|2 represents 1.2E6

     11       1|22222222222
     25       1|77777777777777
     25       2|
     25       2|
     (9)      3|222222222
     22       3|7777
     18       4|
     18       4|
     18       5|22222
     13       5|77777777
      5       6|222
      2       6|77
```

El análisis cuidadoso de este gráfico nos lleva a decidir una estratificación en tres estratos de la siguiente manera:

Estrato I: empresas con ventas diarias de 1.200.000 hasta menoresque3.200.000

Estrato II: empresas con ventas diarias de 3.200.000 hasta menores que 5.200.000

Estrato III: empresas con ventas diarias mayores o iguales a 5.200.000

Si analizamos los coeficientes de variación de cada estrato, vemos que la estratificación ha sido adecuada; en efecto, los respectivos coeficientes de variación son los siguientes:

Summary Statistics

Code	Count	Coefficient of variation
1	25	16.5563%
2	9	0.0%
3	22	16.9364%

Veamos el segundo método para la formación de los estratos. Método acumulativo de la raíz cuadrada de la frecuencia absoluta. Es una técnica adecuada para formar los estratos; veamos como funciona este método con la misma información del ejemplo 2.

De acuerdo a sus ingresos diarios las 56 empresas se agruparon en intervalos de la siguiente manera:

Ventas diarias	*Frecuencia absoluta*	*Frecuencia absoluta*	*Acumulado de la raíz cuadrada de la frecuencia absoluta*
1.000.000–1.500.000	11	3.32	3.32
1.500.000–2.000.000	14	3.74	7.06
3.000.000–3.500.000	9	3.00	10.06
3.500.000–4.000.000	4	2.00	12.06
5.000.000–5.500.000	5	2.24	14.30
5.500.000–6.000.000	8	2.83	17.13
6.000.000–6.500.000	3	1.73	18.86
6.500.000–7.000.000	2	1.41	20.27

Si el auditor desea estratificar esta información en tres estratos, cuáles serían los estratos?

El método acumulativo de la raíz cuadrada de la frecuencia absoluta consiste en encontrar intervalos iguales en la columna del acumulado de la raíz cuadrada de la frecuencia absoluta. Para ello se divide el total de ésta columna por el número de estratos deseados:

Así tenemos, $\frac{20.27}{3} = 6.76$, lo cual nos indica que cada estrato tendría una longitud de 6.76.

Por tanto los límites para los estratos serían:

Para el estrato I: 3.32 a 6.76
Para el estrato II: 6.76 a 13.52
Para el estrato III: 13.52 a 20.27

Como en la escala real en la columna del acumulado de la raíz de la frecuencia absoluta el valor más cercano a 6.76 es 7.06 y el más cercano a 13.52 es 14.30, los tres estratos tendrían las siguientes dimensiones:

Estrato I: empresas con ventas diarias de 1.000.000 a 3.000.000
Estrato II: empresas con ventas diarias de 3.000.001 a 5.500.000
Estrato III: empresas con ventas diarias de 5.500.001 a 7.000.000

8. Muestreo aleatorio por conglomerados. Esta técnica de muestreo se aplica cuando la población se puede clasificar en colecciones de elementos llamados conglomerados con la condición de que estos conglomerados sean similares entre sí. En esta técnica de muestreo podemos dar la siguiente definición:

Definición 14: Una muestra por conglomerados se obtiene seleccionando aleatoriamente un conjunto de n colecciones de la población, llamados conglomerados, y posteriormente se lleva a cabo un censo en cada uno de los conglomerados elegidos.

Observaciones:

- En el muestreo por conglomerados la unidad de muestreo es una colección o conglomerados de elementos.
- Este muestreo es muy utilizado en auditoría, por ejemplo: Para examinar la precisión en la facturación de una compañía en los últimos cinco años, el auditor selecciona aleatoriamente algunos conglomerados que son, por ejemplo, las facturas de un mes y luego lleva a cabo un censo de las facturas de los conglomerados seleccionados. Para verificar la exactitud del procedimiento de pago de una compañía, el auditor selecciona aleatoriamente algunos conglomerados que serían las nóminas semanales, por ejemplo, y luego llevaría a cabo un censo de éstos conglomerados.
- La diferencia entre un estrato y un conglomerado es la siguiente: los estratos deben ser homogéneos pero deben ser diferentes entre sí tanto como sea posible con respecto a la característica que se está estudiando, en cambio, los conglomerados deben ser muy similares entre sí y no se requiere que sean homogéneos en su interior.

Ventajas del muestreo por conglomerados

El muestreo por conglomerados es un diseño efectivo para obtener la máxima información a un costo mínimo bajo las siguientes condiciones: Cuando no se encuentra disponible o es muy costoso obtener un buen marco que liste los elementos de la población y por el contrario es muy fácil obtener un marco que enumere los conglomerados. Cuando la población es grande y está dispersa en una región muy extensa y en consecuencia el costo por obtener las observaciones muestrales se incrementa con la distancia que separa los elementos. En estas dos circunstancias el auditor opta por el muestreo por conglomerados. Veamos un ejemplo. Un auditor desea estimar el ingreso promedio por hogar en una gran ciudad. Si para obtener la muestra utiliza el muestreo aleatorio simple se requiere un marco que liste todos los hogares el cual es muy costoso o imposible de obtener. Si utiliza el muestreo aleatorio estratificado, el costo elevado persiste ya que se necesitaría un marco para cada estrato en la población. Para evitar lo anterior, el auditor puede dividir la ciudad en regiones tales como manzanas, las cuales son los conglomerados, y seleccionar una muestra aleatoria simple entre ellos, y luego medir el ingreso de cada familia en cada manzana seleccionada.

Inconvenientes del muestreo por conglomerados

Generalmente la disminución de costos que produce este muestreo tiene su precio; en efecto: El auditor por conseguir un buen nivel de confianza en la estimación de los parámetros poblacionales debe seleccionar una muestra mucho más grande que la que se emplearía en el muestreo aleatorio simple. Para reducir el error de estimación de los parámetros poblacionales es necesario seleccionar un mayor número de conglomerados pequeños en lugar de pocos conglomerados grandes. El error de estimación puede aumentar debido a que los elementos de un conglomerado tienden a tener características comunes y por tanto se corre el riesgo de que algunas características de la población estén representadas en exceso en un conglomerado en tanto que otras no lo estén. En efecto, la semejanza entre las unidades de un mismo conglomerado puede incrementar el error de estimación ya que la mayoría de la información muestreada es redundante.

Cómo seleccionar la muestra en el muestreo por conglomerados

La primera tarea del auditor en éste diseño muestral es especificar muy bien los conglomerados apropiados; luego debe conformar un marco que liste todos los conglomerados de la población y de éste marco, los conglomerados que van a constituir la muestra los seleccionan con el muestreo aleatorio simple.

9. Muestreo aleatorio por conglomerados en dos etapas

Frecuentemente un conglomerado contiene demasiados elementos lo cual hace difícil su medición total o a veces sus elementos son tan semejantes que la medición de solo unas cuantas unidades proporciona una información completa del conglomerado. Cuando cualquiera de estas dos situaciones se presentan, el auditor puede seleccionar una muestra aleatoria simple de los conglomerados y luego tomar una muestra aleatoria de los elementos de cada conglomerado seleccionado. El resultado es una muestra por conglomerados en dos etapas. En esta técnica de muestreo podemos dar la siguiente definición.

Definición 15: Una muestra aleatoria por conglomerados en dos etapas se obtiene seleccionando primero una muestra aleatoria de conglomerados y posteriormente extrayendo una muestra aleatoria de los elementos de cada conglomerado seleccionado.

Un ejemplo sería el siguiente: una firma posee una cadena de almacenes y el auditor va a estimar el total de cuentas por pagar de la siguiente manera: del total de almacenes de la cadena se eligen n por muestreo aleatorio simple y luego de cada almacén seleccionado, se toma una muestra aleatoria simple del total de sus cuentas por pagar. Cada almacén de la cadena proporciona un conglomerado de cuentas por pagar.

Cómo seleccionar la muestra en el muestreo por conglomerados en dos etapas

Para seleccionar la muestra se tienen en cuenta los siguientes pasos:

1. Se obtiene un marco que liste todos los conglomerados de la población.
2. Aplicando el muestreo aleatorio simple se selecciona una muestra de conglomerados.
3. Se obtienen marcos que enumeren todas las unidades de cada conglomerado seleccionado.

4. Finalmente se selecciona una muestra aleatoria simple o sistemática de los elementos de cada uno de estos marcos si son homogéneos y si no lo son la muestra aleatoria se selecciona en el muestreo estratificado.

Observaciones

- Este diseño muestral elimina la necesidad de muestrear todos los elementos de cada conglomerado elegido, lo cual reduce el costo del muestreo con poca pérdida de información.
- Los conglomerados grandes tienden a contener elementos heterogéneos lo cual hace necesario una muestra grande de cada uno para lograr una estimación precisa de los parámetros poblacionales. En cambio, los conglomerados pequeños frecuentemente contienen elementos relativamente homogéneos, en cuyo caso puede obtenerse una información precisa sobre las características de la población seleccionando una muestra pequeña en cada conglomerado; por tanto el auditor, de acuerdo a los costos, de acuerdo a la naturaleza de la población y de acuerdo a su juicio debe decidir si requiere muestrear pocos conglomerados grandes o muchos conglomerados pequeños.

10. Muestreo con inclusión forzosa

En un proceso de control selectivo el auditor puede encontrar en la población elementos que por su valor cuantitativo elevado o por su importancia cualitativa relevante deben ser incluidos forzosamente en la muestra y no correr el riesgo de que en un proceso estrictamente aleatorio no aparezcan en la muestra seleccionada. En consecuencia estas unidades de observación no deben formar parte del marco muestral. La inclusión en la muestra de estas unidades de observación puede ser de dos formas: Inclusión forzosa imperativa. Inclusión forzosa a criterio del auditor.

Veamos algunos ejemplos de estas dos modalidades de inclusión.

Inclusión forzosa imperativa

Algunos ejemplos de auditoría con inclusión forzosa imperativa en la muestra son:

- *En la revisión de ingresos en forma selectiva en las administraciones de las rentas departamentales.* En este marco muestral de legajos de ingresos se aplica el diseño muestral sistemático o el muestreo aleatorio simple para extraer la muestra, pero si los ingresos están ordenados de acuerdo a su valor se aplicará el muestreo sistemático. El auditor debe hacer inclusión forzosa en la muestra de aquellos recibos con un valor superior a determinada cantidad.

- *En el control selectivo de los recibos de ingresos en la cuenta de tesorería de un instituto.* En el marco muestral del conjunto de recibos de ingresos para la extracción de la muestra se aplica el muestreo aleatorio estratificado. El auditor debe hacer inclusión forzosa en la muestra de los recibos que forman el estrato más alto.
- *En el control selectivo de la caja menor o fondo especial de una entidad.* Para la extracción de la muestra se aplica el muestreo aleatorio por conglomerados. El auditor por indicios, denuncias o por quejas debe hacer inclusión forzosa en la muestra de las cajas menores o fondos especiales cuestionados.
- *En el control selectivo de los vehículos de una entidad.* En este marco muestral del conjunto de vehículos para la extracción de la muestra se aplica el muestreo aleatorio simple. El auditor debe hacer inclusión forzosa de la maquinaria pesada y de los vehículos livianos de alto valor.

Inclusión forzosa a criterio del auditor

Algunos ejemplos de auditorías con inclusión forzosa en la muestra a criterio del auditor son:

- *En la revisión en forma selectiva de los contratos que ha realizado una entidad.* En este marco muestral del conjunto de contratos, para extraer la muestra se aplica el muestreo aleatorio simple. El auditor puede hacer inclusión forzosa de aquellos contratos que según su criterio deben ser revisados.
- *En la revisión en forma selectiva del estado de la cuenta de cada cliente deudor de una entidad.* En este marco muestral del conjunto de clientes deudores se aplica el muestreo aleatorio simple para extraer la muestra si los deudores son personas y se aplica el muestreo por conglomerados en dos etapas si los deudores son entidades. A criterio del auditor

puede incluir en la muestra aquellas cuentas por cobrar que por sus valores elevados son muy representativas.

Elección del método de selectividad para la extracción de la muestra

Una de las preguntas fundamentales del muestreo es: cuál diseño muestral debe seleccionar el auditor para extraer la muestra en cada situación específica de control? Veamos la respuesta a esta pregunta.

1. Elegir el muestreo de criterio, discrecional o intencional. Este método de selectividad lo debe elegir el auditor en las siguientes circunstancias:
   - Si la población es muy pequeña.
   - Si el procedimiento de control es de poca significación o no requiere de proyección estadística.
   - Si la experiencia indica la necesidad de elegir muchos documentos contables dudosos.
2. Elegir el muestreo estadístico. Este método de selectividad aleatorio lo debe elegir el auditor en las siguientes circunstancias:
   - Si la población es mediana o grande.
   - Si se requiere inferencia estadística.
   - Si existe en los funcionarios la preparación y la capacitación suficientes para aplicarlo bien.

La pregunta ahora es: *en cada situación específica de control, cuál diseño muestral estadístico debe elegir el auditor?* Veamos la respuesta a esta pregunta. En cada situación específica de control la elección del diseño muestralestadístico para la extracción de la muestra está determinada por los siguientes factores:

- Por la elección de la variable relevante, la cual según el procedimiento de control es definida como la más importante entre todas las variables que deben ser auditadas.
- Por la disposición de la información en el marco muestral.

- Por la experiencia y conocimiento que tiene el auditor sobre la población y sobre la forma de aplicar el control.

Veamos algunos ejemplos de auditorías específicas, donde se analiza la elección del diseño muestral para la extracción de la muestra.

- *En el control selectivo de los contratos de una entidad.* En esta auditoría, la unidad de observación es el texto del contrato y el marco muestral es el conjunto de contratos. En este procedimiento de control las variables que deben ser controladas son: Variable uno: los requisitos legales. Variable dos: los aspectos técnicos. Variable tres: los aspectos económicos.

La variable relevante es la *uno* la cual es cualitativa; por tanto, para extraer la muestra se debe elegir el muestreo aleatorio simple con inclusión forzosa de algunos contratos a criterio del auditor. El marco muestral debe ser enumerado para la extracción de la muestra.

- *En el control selectivo de órdenes de pago.* La unidad de observación son las ordenes de pago y el marco muestral es el legajo de egresos en la cuenta de tesorería. En este procedimiento de control las variables que deben ser auditadas son: Variable uno: la correcta liquidación. Variable dos: la correcta imputación presupuestal. Variable tres: los soportes del pago. Variable cuatro: firma del ordenador y del pagador. Variable cinco: firma de recibo. Variable seis: otros requisitos según el manual de procedimientos de la entidad.

La variable relevante es la *uno* la cual es cuantitativa y por tanto, para extraer la muestra se debe aplicar el muestreo aleatorio estratificado con inclusión forzosa de las órdenes de pago del estrato más alto y de las ordenes de pago derivadas de contratos por licitación o por adjudicación directa. Para la extracción de la muestra se deben enumerar los recibos de egresos de cada estrato.

- *En la conciliación de la cuenta bancaria de forma selectiva.* La unidad de observación es la conciliación de la cuenta bancaria y el marco muestral es el conjunto de conciliaciones en la cuenta de tesorería. En este procedimiento las variables que deben ser auditadas son: Variable uno: correcta conciliación. Variable dos: los procedimientos señalados en el manual de auditoría.

La variable relevante es la *uno* la cual es cualitativa; por tanto para la extracción de la muestra se ha de elegir el muestreo aleatorio por conglomerados. El proceso de conciliación se hace de forma total en aquellas cuentas bancarias que han sido seleccionadas. Para la selección de la muestra se enumeran las cuentas bancarias que forman el marco muestral. Es de advertir que sólo se hará conciliación de forma selectiva si el conjunto de cuentas bancarias es superior a un número predeterminado.

- *En el control selectivo de los bienes devolutivos dentro de las mínimas unidades administrativas o del listado de empleados.* La unidad de observación son los bienes devolutivos y el marco muestral es el listado de los bienes. En este procedimiento las variables que deben ser auditadas son: Variable uno: la existencia del bien. Variable dos: la verificación del código correcto. Variable tres: la verificación de la descripción correcta del bien y de su placa. Variable cuatro: su valor adecuado. Variable cinco: verificar que el inventario se haya rendido en los formatos exigidos.

La variable relevante es la *uno* y es cualitativa. Por tanto, para la extracción de la muestra se ha de elegir el muestreo aleatorio por conglomerados. La selectividad se hace sobre las mínimas unidades administrativas o sobre el listado de bienes del empleado. Una vez seleccionadas, si los bienes son menos de una determinada cantidad se hace inspección a todos los bienes y si el número de artículos supera la anterior cantidad se hace una nueva selección aplicando el muestreo aleatorio simple. Como se ve, en este segundo caso se aplica el muestreo por conglomerados en dos etapas. En este control de los bienes devolutivos, el auditor debe hacer inclusión forzosa en la muestra de aquellos bienes con alto valor artístico, cultural y monetario.

- *En la revisión de ingresos en forma selectiva en las administraciones de rentas departamentales.* La unidad de observación son los recibos de ingreso y el marco muestral es el legajo de ingresos en la cuenta de tesorería. En este procedimiento las variables que deben ser auditadas son: Variable uno: la correcta liquidación del valor. Variable dos: verificar la fecha. Variable tres: verificar la firma del cajero. Variable cuatro: verificar que se cumplan los otros procedimientos que están en el manual de auditoria. La variable relevante en este proceso de control es la *uno* la cual es cualitativa y para la extracción de la muestra se ha de elegir el muestreo aleatorio sistemático. Además, se ha de hacer inclusión forzosa de los recibos superiores a una cantidad determinada. Se ha de anotar que la verificación del número consecutivo de los recibos de los ingresos no admite selectividad.

- *En el control selectivo de los ingresos de la dirección de tránsito.* La unidad de observación es cada factura y el marco muestral es el legajo de ingresos. En este procedimiento las variables que deben ser auditadas son: Variable uno: la correcta liquidación. Variable dos: la correcta imputación presupuestal. Variable tres: la revisión de otros aspectos que el auditor considere necesarios.

En este proceso de control la variable relevante es la *uno* la cual es cualitativa y para la extracción de la muestra se ha de elegir el muestreo sistemático especial ya que no se conoce el valor de N necesario para calcular el valor de k.

Nota: Con el estudio de la elección del diseño muestral para la extracción de la muestra, hemos dado respuesta a la primera gran pregunta del muestreo: *cómo se extrae la muestra de una población.*

III. Tamaño de la muestra en los diferentes diseños muestrales

Vamos a darle respuesta a la segunda gran pregunta del muestreo: *cuál debe ser el tamaño de la muestra.* En efecto, uno de los principales problemas que debe solucionar el auditor al principio de toda investigación por muestreo es el tamaño de muestra a seleccionar. El cálculo inicial del tamaño de la muestra dependerá básicamente de dos factores:

- De la precisión que se desea en la estimación de los parámetros poblacionales.
- De la confiabilidad con la cual se desea llevar a cabo tales estimaciones.

En efecto, la precisión no es otra cosa que el valor del error máximo absoluto que se puede tolerar en las estimaciones muestrales; es decir, la precisión es la diferencia máxima, que denotaremos por E, que puede aceptarse entre la estimación puntual del parámetro poblacional y el verdadero valor del parámetro; por tanto, E es la cantidad del error que puede tolerarse al estimar el parámetro poblacional. Su definición es:

$$\left| valor\ del\ estimador - valor\ verdadero \right| \leq E$$

En la determinación del tamaño de la muestra otro factor determinante es la confiabilidad con la cual se desea trabajar. Algunas veces se habla de riesgo en vez de confiabilidad. Este riesgo es el mismo valor utilizado en el cálculo de los intervalos de confianza y es la probabilidad de que el intervalo de confianza calculado no contenga el verdadero parámetro que se desea estimar.

Otras consideraciones que el auditor debe tener en cuenta en el cálculo del tamaño de la muestra son:

- Si existe un desconocimiento total de las características de la población, el tamaño de la muestra se calcula con la fórmula del diseño muestral aleatorio de población infinita o finita.
- Si existen varias variables de interés, el auditor debe decidir con anterioridad cuál es la variable más importante y en base a esta variable relevante se calcula el tamaño de la muestra. Así, en un control de las ventas diarias de un almacén, por ejemplo, se tienen características cuantitativas y características cualitativas; el auditor decide cuál es la variable más importante y de acuerdo a esta variable relevante determina el tamaño de la muestra.
- Si se calcula el tamaño de la muestra en base a una variable cuantitativa y se hace también en base a una variable cualitativa de la información, se obtendrán diferentes tamaños de muestra. En esta situación si los recursos lo permiten, se debe seleccionar el máximo valor de muestra encontrado. En caso contrario lo recomendable es hacer un promedio de los dos valores encontrados.
- Es necesario cuantificar los errores no muestrales para introducir los correctivos en el tamaño de la muestra. Errores no muestrales son, por ejemplo: Errores en la información de la población. Errores en el procesamiento de la información. No respuestas. Respuestas incompletas. Encuestas incompletas o mal diligenciadas.

Una vez detectados adecuadamente los errores no muestrales la muestra debe ser ajustada por el siguiente factor de ajuste: $\frac{1}{1-D}$, donde D es la proporción total estimada de errores no muestrales. Así, por ejemplo, si basados en análisis anteriores similares, se estima que el porcentaje de no respuestas en una encuesta es del 15% y que el porcentaje de encuestas inconsistentes puede alcanzar un 10%, el factor de ajuste sería:

$$n = n_1 * \frac{1}{1-0.25}$$

donde n_1 es el tamaño de la muestra obtenido de acuerdo a las formulas correspondientes.

Vamos a estudiar la determinación del tamaño de la muestra en los diferentes diseños muestrales.

1. Tamaño de la muestra en el diseño muestral aleatorio

1.1. Tamaño de la muestra para la estimación de la media y del total poblacional cuando la variable relevante es cuantitativa

1.1.1. En muestreo de poblaciones infinitas

Si observamos el intervalo de confianza para μ cuando se trabaja con la distribución normal con σ conocido, el límite de precisión es:

$$Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

el cual es el margen de error en la estimación, luego:

$$E = Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

De esta desigualdad despejando a n, podemos obtener el valor del tamaño de la muestra que denotaremos por n_∞, así:

$$n_\infty = \frac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * \sigma^2}{E^2}$$

Observaciones

- El auditor decide el nivel de confianza con el cual va a trabajar lo que determina el valor del percentil $Z_{1-\frac{\alpha}{2}}$, hallado en la distribución normal estandarizada. El auditor para asignar el valor de $Z_{1-\frac{\alpha}{2}}$ lo hará de acuerdo a la evaluación que haga del control interno de la entidad que controla; para ello debe evaluar tres aspectos: Que exista el control interno. Que esté bien diseñado. Que se aplique correctamente.

De acuerdo a la evaluación de la calidad y de la eficacia del control interno, el auditor define el nivel de confianza el cual incide en el tamaño de la muestra. Lógicamente la relación entre el control interno y el nivel de confianza es inversamente proporcional. Así, si el control interno es bueno el tamaño de la muestra es más bajo que si el control interno es deficiente. En consecuencia, si el control interno es deficiente el auditor exige una confianza en la prueba mayor lo cual da un valor de $Z_{1-\frac{\alpha}{2}}$ alto, lo cual conlleva a un tamaño mayor de la muestra.

Como ejemplo, el auditor podría adoptar la siguiente regla de decisión:

Evaluación del control interno	Confianza en la prueba	Valor de $Z_{1-\frac{\alpha}{2}}$ en la distribución normal
Bueno	Baja 90%	1.645
Regular	Media 95%	1.96
Deficiente	Alta 99%	2.575

- El auditor elige el máximo límite de precisión que puede tolerar. La revisión de estudios similares anteriores y la experiencia y buen juicio del auditor son factores decisivos en la determinación del error.
- El auditor debe conocer el valor de la varianza poblacional σ^2, en caso contrario es necesario estimarla con la varianza muestral S^2; pero aquí

aparece un circulo vicioso ya que para estimar a es necesario elegir una muestra n y para elegir a n se necesita conocer de antemano a . Para romper éste circulo vicioso se procede así: El auditor estima a de acuerdo al conocimiento en trabajos similares en otras empresas o en períodos anteriores en la misma empresa. La varianza más aceptable encontrada en estas revisiones se ha de tomar como la estimación de las varianza real . En todo caso, es preferible sobrestimar la dispersión para no subvalorar el tamaño de la muestra. El buen juicio del auditor juega un papel importantísimo en la decisión de la estimación de la varianza. El auditor estima a tomando la muestra en dos etapas. El auditor inicia el proceso tomando una muestra inicial de tamaño la cual es llamada muestra piloto; calcula en esta muestra la desviación típica de la característica que está estudiando y la utiliza como estimador de para calcular el tamaño de muestra definitivo.

Cuando se toma una muestra preliminar, el tamaño de muestra debe ser ajustado por el factor $\left(1+\frac{2}{n_1}\right)$ que es el precio a pagar por el no conocimiento de , donde es el tamaño de la muestra piloto. El tamaño de la muestra definitiva es:

$$n_{\infty} = \frac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * S^2}{E^2}\left(1+\frac{2}{n_1}\right)$$

En algunos casos el auditor puede tener una idea bastante aproximada de a distribución de la población y en estas circunstancias es posible encontrar una buena aproximación de la varianza poblacional. Si la distribución de la población en estudio se puede asumir como normal y se pueden encontrar valores aproximados de la observación mínima y máxima, el rango se puede utilizar como estimador de la varianza. En efecto, sabemos que en la distribución normal el 99.7% de las observaciones caen en el intervalo:

$$\left[media - 3\ desviaciones \quad , \quad media + 3\ desviaciones\ \right]$$

En consecuencia, la longitud de este intervalo puede hacerse igual al rango, por lo cual tendríamos: $Rango = 2 * 3 * \sigma$, de donde:

$\sigma^2 = \left(\frac{Rango}{6} \right)^2$, y este valor sería el estimador de σ^2.

- El tamaño de la muestra es directamente proporcional a la confiabilidad deseada y a la varianza de la población. Esto significa que un aumento en la confiabilidad conlleva a un aumento en el tamaño de la muestra, lo mismo que un aumento en la dispersión poblacional se traduce en un aumento en el tamaño muestral.
- El tamaño de la muestra es inversamente proporcional al valor del error elegido por el auditor. La sensibilidad del tamaño de la muestra respecto a pequeñas variaciones en el error es manifiesta. En efecto, una disminución en el error tiene el precio de un gran aumento en el tamaño de la muestra.

1.1.2. En muestreo de poblaciones finitas

Cuando se trabaja con la distribución normal con conocido, el límite de precisión en los intervalos de confianza para es:

$$Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$$

el cual es el margen de error en la estimación; luego,

$$E = Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} * \sqrt{\frac{N-n}{N-1}}$$

De esta relación con un poco de álgebra despejando a n se obtiene el valor del tamaño de la muestra que denotaremos por , así:

$$n_c = \frac{N * \left(Z_{1-\frac{\alpha}{2}} \right)^2 * \sigma^2}{N * E^2 + \left(Z_{1-\frac{\alpha}{2}} \right)^2 * \sigma^2}$$

Esta formula la podemos escribir así:

$$n_c = \frac{\dfrac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * \sigma^2}{E^2}}{1 + \dfrac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * \sigma^2}{N * E^2}}$$

Esta formula, se suele expresar en función de , de la siguiente manera:

$$n_c = \frac{n_\infty}{1 + \dfrac{n_\infty}{N}}$$

Observaciones

- Esta formula es función del percentil de la distribución normal estándar, del error, de la varianza y del tamaño poblacional *N*.
- Anteriormente en la formula de se analizó la relación de las variables $Z_{1-\frac{\alpha}{2}}$, E y , E y con el tamaño de la muestra. La incidencia de la otra variable de la formula, *N*, en el tamaño muestral es así: un aumento en *N* produce un aumento de diferente intensidad en *n*. En efecto, para valores pequeños de *N*, el valor de *n* se verá sensiblemente afectado por cambios en *N*; pero si el valor de *N* es grande (alrededor de 10.000), el valor de *n* es prácticamente insensible ante cambios en *N*.
- Siempre $n_c < n_\infty$. Por tanto $n_{\infty.} = \dfrac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * S^2}{E^2}$ es la cota superior de n_c.

  En consecuencia, un tamaño de muestra superior a sería un desperdicio innecesario de recursos.

Nota. En la práctica para calcular el tamaño de la muestra para estimar la media y el total poblacional se procede de la siguiente manera: La muestra se extrae sin sustitución de elementos. Se calcula el valor de n_{∞}. Se calcula el valor del factor de corrección $\frac{N-n}{N-1}$ tomando como valor de n el calculado en . Si $\frac{N-n_{\infty}}{N-1} \geq 0.9$, el tamaño de la muestra es el encontrado en y si el valor del factor de corrección es $\frac{N-n_{\infty}}{N-1} < 0.95$, la muestra se puede corregir disminuyéndola un poco y el valor del tamaño de la muestra es .

Lo anterior equivale a lo siguiente: si ocurre que $n_{\infty} \leq 0.05 * N$ la muestra es pequeña por lo cual no se puede disminuir y en consecuencia el tamaño de la muestra es , pero si ocurre que $n_{\infty} > 0.05 * N$, la muestra se considera grande y puede ser corregida disminuyéndola un poco calculando a que sería el tamaño definitivo de la muestra.

Todas las consideraciones anteriores acerca del cálculo del tamaño de la muestra lo vamos a aplicar en el siguiente ejemplo.

Ejemplo 1. Una compañía que posee una cadena de almacenes ha repartido entre sus clientes especiales tarjetas de crédito a fin de promover sus ventas. El auditor de esta compañía quiere hacer un estudio sobre la utilización de dicha tarjeta de crédito. Para ello, desea en un mes de referencia, determinar el tamaño de la muestra para estimar con un error del 5% y una confianza del 95%, el gasto promedio mensual cancelado con la tarjeta de crédito. De acuerdo a su experiencia el auditor estima que el gasto promedio mensual cancelado con estas tarjetas es de 100.000 pesos con una dispersión de 30.000 pesos.

Solución 1. Como la variable relevante es cuantitativa, para hallar el tamaño de la muestra se tienen los siguientes valores:

- Como el auditor ha elegido una confianza del 95%, por la simetría de la distribución normal, es necesario hallar en la normal estandarizada el percentil 97.5 cuyo valor es: $Z_{0.975} = 1.96$

- $\sigma^2 = (30.000)^2$

- $E = 5000$

ya que el auditor ha decidido tolerar una diferencia máxima del 5% entre el estimador de la media y su verdadero valor. Luego:

$$n_{\infty} = \frac{(1.96)^2 * (30.000)^2}{(5.000)^2} = 138.2976$$

Si el valor del tamaño muestral es un decimal, este debe aproximarse al entero siguiente ya que de otra forma se disminuye la precisión y confiabilidad elegidas inicialmente; por tanto, $n_{\infty} = 139$.

En consecuencia, para estimar el consumo promedio mensual pagado con tarjeta de crédito, con una confianza del 95% y un error del 5%, es necesario analizar el pago del consumo mensual de 139 clientes poseedores de tarjeta de crédito.

En este problema haremos las siguientes variaciones:

Variación uno: cambiar el error del 5% por un error del 2% y del 1%.

Con un error del 2%, se tiene:

$$n_{\infty} = \frac{(1.96)^2 * (30.000)^2}{(2.000)^2} = 865$$

Con un error del 1%, se tiene:

$$n_{\infty} = \frac{(1.96)^2 * (30.000)^2}{(1.000)^2} = 3458$$

Se puede observar la gran sensibilidad de la muestra ante cambios en la variable error. En la práctica, cuando es necesario aumentar el tamaño de la muestra, se hace reduciendo la variable error.

Variación dos: resolver el problema inicial con una confianza del 99%. Sabemos que $P_{0.995} = 2.585$, hallado en la *n(0, 1)*. De modo que:

$$n_{\infty} = \frac{(2.585)^2 * (30.000)^2}{(5.000)^2} = 241$$

Se observa que una mayor confiabilidad, exige un aumento en el tamaño de la muestra.

Variación tres: la dispersión estimada en los gastos promedios mensuales cancelados con tarjeta de crédito es de 32.000 pesos.

$$n_{\infty} = \frac{(1.96)^2 * (32.000)^2}{(5.000)^2} = 158$$

Hay un aumento en el tamaño de la muestra ya que ésta es directamente proporcional al valor de la varianza de la población.

Variación cuatro: supongamos que en la recolección de la información se presentó la siguiente anomalía: no respuestas un 12% y un 9% de respuestas inconsistentes, es decir, se presentó un error no muestral cuya cuantificación es del 21%. Por tanto, el tamaño de la muestra original debe ser ajustado por el factor $\frac{1}{1-0.21}$.

En consecuencia:

$$n_{\infty} = \frac{(1.96)^2 * (30.000)^2}{(5.000)^2} * \frac{1}{1-0.21} = 176$$

Para compensar los errores no muestrales cometidos en la recolección de la muestra, ésta debe ser aumentada de 139 a 176 usuarios de la tarjeta de crédito.

Variación cinco: en el problema original, no se sabe nada de las características de la población, es decir, ni se conoce su media ni su dispersión. Por tanto, para estimarlos el auditor toma la muestra en dos etapas: primero toma una muestra piloto, con la cual calcula los estimadores de los parámetros poblacionales y luego, con éstos resultados calcula el tamaño de muestra definitivo.

Supongamos que el auditor tomó una muestra preliminar de tamaño 30, donde encontró que el promedio de los gastos mensuales cancelados con la tarjeta de crédito en el mes en referencia fueron de 105.000 pesos con una dispersión de 32.000 pesos.

En esta situación tenemos los siguientes valores:

$$P_{0.975} = 1.96$$
$$S^2 = (32.000)^2$$
$$E = 5250$$

El tamaño de la muestra debe ser ajustado por el factor $1 + \frac{2}{30}$ que es el precio de la estimación de la varianza poblacional.

$$n_{\infty} = \frac{(1.96)^2 * (32.000)^2}{(5.250)^2} * \left(1 + \frac{2}{30}\right) = 153$$

Por tanto, para estimar el consumo promedio mensual pagado con tarjeta de crédito, con una confianza del 95% y un error del 5%, es necesario analizar el pago del consumo mensual de 153 clientes poseedores de la tarjeta de crédito.

Variación seis: no se conoce la varianza poblacional pero el auditor estima que casi la totalidad de los clientes tienen un consumo mensual entre 80.000 y 120.000 pesos.

En esta situación puede asumir que cualquier consumo dentro de éste intervalo es igualmente probable y por tanto la distribución de consumo pagado con tarjeta de crédito sigue aproximadamente una distribución uniforme. Además, sabemos que la varianza de esta distribución está dada por $\frac{(b-a)^2}{12}$ donde a y b son los valores mínimo y máximo de la variable. En el ejemplo, y , por tanto:

$$S^2 = \frac{(120.000 - 80.000)^2}{12} = 133'333.333{,}33$$ valor que se toma como el estimador de σ^2.

En consecuencia:

$$n_{\infty} = \frac{(1.96)^2 * (11.547)^2}{(5.000)^2} = 21$$

Se puede observar una gran disminución en el tamaño de la muestra debido a la gran disminución que se tiene en la variabilidad, en efecto, ésta pasa de 30.000 a 11.547 pesos.

Variación siete: las características poblacionales no se conocen pero razonablemente el auditor supone que el pago del consumo mensual con tarjeta de crédito sigue aproximadamente una distribución normal entre un consumo mínimo de 50.000 pesos y un máximo de 150.000. Por tanto, por las propiedades de la distribución normal tenemos:

$$\sigma^2 = \left(\frac{Rango}{6}\right)^2 = \left(\frac{100.000}{6}\right)^2 = 277'777.777,8$$

que es el valor de la varianza estimada. El tamaño de la muestra sería:

$$n_{\infty} = \frac{(1.96)^2 * 277'777.777,8}{(5.000)^2} = 43$$

Como se puede observar, cuando la distribución poblacional es normal, para la estimación de la media y del total, basta con una muestra muy pequeña. Si se quiere aumentar el tamaño de la muestra, trabajaríamos con un error, por ejemplo del 2% y el tamaño muestral sería:

$$n_{\infty} = \frac{(1.96)^2 * 277'777.777,8}{(2.000)^2} = 267$$

Variación ocho: la compañía le dio tarjetas de crédito a 2.000 clientes especiales. Al tener en cuenta la variable N en el cálculo del tamaño de la muestra debemos obtener el valor de

$$n_c = \frac{N * \left(Z_{1-\frac{\alpha}{2}}\right)^2 * \sigma^2}{N * E^2 + \left(Z_{1-\frac{\alpha}{2}}\right)^2 * \sigma^2} = \frac{n_{\infty}}{1 + \frac{n_{\infty}}{N}}$$

Entonces:

$$n_c = \frac{139}{1+\frac{139}{2000}} = 130$$

Como el factor de corrección de población finita $\frac{N-n_\infty}{N-1}$ es menor que 0.95, el tamaño de se puede corregir y el tamaño de la muestra es de 130 consumos mensuales de usuarios con tarjeta de crédito. A esta misma conclusión llegamos al ver que el tamaño de es superior al 5% de N, por tanto se considera una muestra grande que se puede reducir un poco con el cálculo de .

Variación nueve: la compañía le dio tarjetas de crédito a 3.000 clientes.

Como se tiene que $n_\infty \leq 0.05 * N$, la muestra no puede ser corregida ya que se considera pequeña y en consecuencia el tamaño de la muestra es $n_\infty = 139$.

Igual conclusión se obtiene al observar el factor de corrección $\frac{N-n_\infty}{N-1}$ el cual es superior a 0.95.

Variación diez: la compañía le dio tarjetas a 200 clientes, a 500 clientes, a 5.000 clientes, a 10.000 clientes, a 100000 clientes y a 1000000 de clientes.

Los diferentes tamaños muestrales serían:

N=200	$n_c = 82$
N=500	$n_c = 108$
N=3.000	$n_c = 133$
N=5.000	$n_c = 136$
N=10.000	$n_c = 137$
N=100.000	$n_c = 138$
N=1.000.000	$n_c = 139$

Se observa que un aumento en *N* produce un aumento de diferente intensidad en *n*. Para valores pequeños de la población, el tamaño de la muestra es afectado sensiblemente por cambios en *N*; pero si el valor de *N* es grande, el valor del tamaño de la muestra es prácticamente insensible ante cambios en el tamaño de la población. En ésta propiedad radica una de las grandes ventajas del muestreo. En efecto, si el tamaño de la población es grande, alrededor de 10.000, el tamaño de la muestra es igual para una población de 20.000 o 200.000 o 2.000.000 o más.

Aplicación Informática

En el paquete estadístico podemos hallar el tamaño de la muestra para la estimación de la media y del total poblacional; hallémosla para la información del ejemplo 1. Los comandos son:

Se parte del comando DESCRIBE, luego se elige SAMPLE SIZED DETERMINATION y se selecciona NORMAL MEAN; en HYPOTHESIZED MEAN se escribe el valor del promedio estimado, en nuestro ejemplo se escribe 100.000 y en HYPOTHESIZED SIGMA se escribe el valor de la dispersión estimada, en el ejemplo 30.000; luego se escribe la cantidad del error elegido por el auditor en ABSOLUTE ERROR, pero es mejor utilizar la opción de escribir el error en porcentaje en vez de calcularlo en valor absoluto y lo escribimos en RELATIVE ERROR en nuestro ejemplo 5%. En CONFIDENCE LEVEL se selecciona el porcentaje de confianza para la estimación que ha elegido el auditor; con clic derecho en ANALYSIS OPTIONS se puede cambiar el error en valor absoluto o mejor en valor relativo.

La salida del computador:

Sample-Size Determination

Parameter to be estimated: normal mean

Desired tolerance: +- 5.0% when mean = 100000.0

Confidence level: 95.0%

Assumed sigma: 30000.0

The required sample size is n=141 observations.

1.2. Tamaño de la muestra para la estimación de la proporción poblacional cuando la variable relevante es cualitativa

1.2.1. En muestreo de poblaciones infinitas

Si observamos el intervalo de confianza para p cuando se trabaja con la distribución aproximadamente normal con conocido vemos que el límite de precisión es:

$$Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}}$$

luego

$$E = Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}}$$

De esta igualdad despejando a n podemos obtener el valor del tamaño de la muestra que denotaremos por , así:

$$n_{\infty} = \frac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * p * (1-p)}{E^2}$$

Todos los términos de la formula han sido definidos anteriormente y el análisis es similar al que se hizo para en la estimación de la media y del total poblacional. En general la proporción poblacional, p, que determina la varianza es desconocida y debe ser estimada previamente para determinar el valor de n. Para su estimación el auditor puede proceder de acuerdo a las siguientes opciones:

- Utilizar el conocimiento que tenga sobre el problema en cuestión.
- Tomar una muestra inicial piloto de 100 ó 200 unidades, las cual le proporciona una primera estimación de p y así calcular el valor de n. En este caso el tamaño de la muestra debe ser ajustado por el factor $\left(1+\frac{2}{n_1}\right)$,

donde es el tamaño de la muestra piloto.

- Darle el auditor a p el valor de $\frac{1}{2}$, con lo cual se obtiene el máximo valor que n puede tomar; un valor diferente a $p = \frac{1}{2}$ produce un valor de tamaño de muestra menor debido a que la varianza $\sigma^2 = p(1-p)$ es creciente para valores de p entre 0 y 0.50; además sabemos que mientras más grande sea la varianza mayor será el tamaño de la muestra y más segura la estimación.
- En algunas circunstancias el auditor puede establecer son anterioridad unos límites para el valor de la proporción poblacional, p, que se quiere estimar. En este caso se elegirá como estimador de p el valor más cercano a $\frac{1}{2}$ entre todos los valores posibles que contiene el límite especificado para p. Por ejemplo, si el auditor sabe que en cierta compañía el porcentaje de facturas con error es como máximo el 18%, se tomaría como estimador de p el 18%. Ahora, si el auditor sabe, en un estudio de costos, que el porcentaje de artículos cuyo costo es superior a 50.000 pesos es al menos del 70%, la estimación de p sería 0.70.

1.2.2. En muestreo de poblaciones finitas

Cuando se trabaja con la distribución normal con conocido, el límite de precisión en los intervalos de confianza para p, es:

$$Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}} * \sqrt{\frac{N-n}{N-1}}$$

luego:

$$E = Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}} * \sqrt{\frac{N-n}{N-1}}$$

De esta igualdad con un poco de álgebra despejando a n se obtiene el valor del tamaño de la muestra que denotaremos por así:

$$n_c = \frac{N * \left(Z_{1-\frac{\alpha}{2}}\right)^2 * p * (1-p)}{N * E^2 + \left(Z_{1-\frac{\alpha}{2}}\right)^2 * p * (1-p)}$$

Esta formula la podemos expresar así:

$$n_c = \frac{\dfrac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * p * (1-p)}{E^2}}{1 + \dfrac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * p * (1-p)}{N * E^2}}$$

y se suele expresar en función de de la siguiente manera:

$$n_c = \frac{n_\infty}{1 + \dfrac{n_\infty}{N}}$$

El análisis de esta formula es similar al que se hizo para en la estimación de la media y del total poblacional.

En la práctica para calcular el tamaño de la muestra para estimar a p se procede de la siguiente manera:

1. Se calcula primero el tamaño de la muestra en población infinita, , que da el mayor tamaño muestral.
2. Se calcula con el factor de corrección TAB 68. Si el factor de corrección es al menos 0.95, el tamaño de la muestra es . Ahora, si $\frac{N - n_\infty}{N-1} < 0.95$,

el tamaño de la muestra es n_c.

La misma decisión se toma al analizar el tamaño de con relación a N. Es decir, si , el tamaño de la muestra es y si $n_\infty \leq 0.05 * N$,el tamaño de la muestra es n_∞ y si $n_\infty > 0.05 * N$,ya que la muestra puede ser corregida disminuyéndola un poco.

Apliquemos las anteriores anotaciones en el siguiente ejemplo:

Ejemplo 2. La compañía de la cadena de almacenes del problema 1 quiere en el mes de referencia, determinar el tamaño de la muestra para estimar con un error del 5% y una confianza del 95%, el porcentaje de clientes que cancelan sus compras mensuales con la tarjeta de crédito.

Solución 2. Como no se conoce el valor de p y el auditor quiere obtener el máximo tamaño de muestra posible para asegurar una buena estimación de la proporción poblacional p, para ello el auditor le da a p el valor $\frac{1}{2}$.

Los valores de las otras variables que entran en la fórmula los conocemos, en efecto:

$$P_{0.975} = 1.96$$
$$E = 5\%$$

De este modo, el tamaño de la muestra es:

$$n_\infty = \frac{(1.96)^2 * 05 * 05}{0.05^2} = 385$$

El auditor eligió un error del 5% y una confianza del 95%; para estos valores 385 es el valor máximo que la muestra puede tomar. En consecuencia, para estimar el porcentaje de tenedores de la tarjeta de crédito que pagan sus compras mensuales con la tarjeta es necesario analizar 385 clientes.

Variación uno: el auditor para estimar el valor de p tomó una muestra piloto de 100 clientes donde encontró que el 70% de los clientes con tarjeta la utilizan para pagar sus compras mensuales.

En este caso, el tamaño de la muestra es:

$$n_{\infty} = \frac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * p * (1-p)}{E^2} * \left(1 + \frac{2}{n_1}\right)$$

ya que es necesario el ajuste por estimar a p con una muestra previa inicial.

$$n_{\infty} = \frac{(1.96)^2 * 0.70 * 0.30}{0.05^2} * \left(1 + \frac{2}{100}\right) = 330$$

Variación dos: supongamos que en la recolección de la información se presentó un 10% de no respuestas y un 8% de respuestas inconsistentes, en consecuencia, se presentó un error no muestral de una cuantía del 18%. Por tanto, el tamaño de la muestra original debe ser ajustado por el factor $\frac{1}{1-0.18}$.

Utilizando el estimador de p de la muestra piloto, el tamaño de la muestra es:

$$n_{\infty} = \frac{(1.96)^2 * 0.70 * 0.30}{0.05^2} * \frac{1}{1-0.18} = 394$$

Variación tres: el auditor estima que al menos el 60% de los clientes con tarjeta de crédito pagan sus compras mensuales con la tarjeta de crédito.

En este caso, asumimos como estimador de p el valor 0.60; por tanto el tamaño de la muestra es:

$$n_{\infty} = \frac{(1.96)^2 * 0.60 * 0.40}{0.05^2} = 369$$

Variación cuatro: la compañía le dio la tarjeta de crédito a: i) 10.000 clientes y ii) a 3.000 clientes.

Utilizando la estimación de p de la muestra piloto, el tamaño de la muestra es:

i) Hallamos primero el tamaño de la muestra con .

$$n_{\infty} = \frac{(1.96)^2 * 0.70 * 0.30}{0.05^2} = 323$$

y calculamos luego el factor de corrección $\frac{N - n_{\infty}}{N - 1} = 0.9678$; como el factor de corrección es superior a 0.95, la muestra no debe ser corregida y el tamaño de la muestra es $n_{\infty} = 323$.

Como $n_{\infty} \leq 0.05 * N$, decidimos que la muestra es pequeña por lo cual no debe ser reducida.

ii) Hallemos primero i) n_{∞} , $n_{\infty} = 323$. y calculamos luego el factor de corrección $\frac{N - n_{\infty}}{N - 1} = 0.8976$; como el factor de corrección es menor que 0.95, la muestra puede ser corregida y ser disminuida un poco, así:

$$n_c = \frac{n_{\infty}}{1 + \frac{n_{\infty}}{3000}} = 292$$

Como , la corrección de está plenamente justificada. El tamaño de la muestra se puede reducir de 323 a 292.

Aplicación Informática

En el paquete estadístico STATGRAPHICS podemos hallar el tamaño de la muestra para estimar la proporción poblacional; los comandos son:

Se elige la opción DESCRIBE, luego SAMPLE DETERMINATION y se selecciona BINOMIAL PROPORTION; en HYPOTHESIZED PROPORTION se registra el valor de la estimación de la proporción poblacional y en RELATIVE ERROR se registra el error en porcentaje elegido por el auditor. Con clic derecho en ANALYSIS OPTIONS se puede cambiar el error.

Antes de hacer un ejemplo analicemos lo siguiente: en el cálculo del tamaño de la muestra para estimar la proporción poblacional, la formula

$$n_{\infty} = \frac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * p * (1-p)}{E^2}$$

Como es basada en la simetría de la distribución normal presenta el siguiente problema: el tamaño de la muestra es igual para $p=0.10$ que para $p=0.90$ lo cual no es lógico. En efecto, si p es muy pequeño se requiere un gran tamaño de muestra para detectar un elemento con la característica de interés. En cambio si p es alto, basta una muestra pequeña para detectar un elemento con la característica de interés. En consecuencia, no es lógico que el tamaño de la muestra para un valor de p bajo sea el mismo que para un valor de p alto simétrico. El paquete estadístico STATGRAPHICS PLUS está diseñado, basado en la distribución binomial, para calcular el tamaño de la muestra en función de la magnitud del valor de p. En efecto, con un margen de error del 5% y una confianza del 95% si $p=0.90$ el tamaño de la muestra es de 242 unidades pero si $p=0.10$ el tamaño de la muestra se eleva a 14.086 unidades; lo cual es lógico, porque si se tiene un p bajo es necesario tener una muestra de gran tamaño para poder detectar un elemento con la características de interés.

Comprobemos lo anterior en el siguiente ejercicio. Con un error del 5% y una confianza del 95%, veamos el cálculo del tamaño de la muestra en el paquete estadístico para los siguientes valores de p: 0.90, 0.50, 0.10, 0.05. La salida del computador es:

Sample-Size Determination

Parameter to be estimated: binomial parameter
Desired tolerance: +- 5.0% when proportion = 0.9
Confidence level: 95.0%

The required sample size is n=242 observations.

Sample-Size Determination

Parameter to be estimated: binomial parameter
Desired tolerance: +- 5.0% when proportion = 0.5
Confidence level: 95.0%

The required sample size is n=1697 observations.

Sample-Size Determination

Parameter to be estimated: binomial parameter
Desired tolerance: +- 5.0% when proportion = 0.1
Confidence level: 95.0%

The required sample size is n=14086 observations.

Sample-Size Determination

Parameter to be estimated: binomial parameter
Desired tolerance: +- 5.0% when proportion = 0.05
Confidence level: 95.0%

The required sample size is n=29858 observations.

Se observa que el tamaño de la muestra no es el mismo para valores simétricos de p, al contrario, cuando el valor de p se reduce el tamaño de la muestra se eleva en forma ostensible.

2. Tamaño de la muestra en el diseño muestral aleatorio estratificado

2.1. Tamaño de la muestra para la estimación de la media y del total poblacional cuando la variable relevante es cuantitativa

En una población estratificada en H estratos, hay que considerar dos aspectos: cuál debe ser el tamaño de la muestra y luego, cómo se reparte la muestra en los diferentes estratos. La repartición de la muestra es llamada afijación o asignación de la muestra. Estos dos aspectos están influenciados por tres factores:

1. El número de elementos en cada estrato.
2. La variabilidad de las observaciones dentro de cada estrato.
3. El costo por obtener una observación de cada estrato.

En efecto, el número de elementos de cada estrato afecta la cantidad de información de la muestra. Lógicamente una muestra de tamaño 30 de una población de 300 elementos contiene más información que una muestra de tamaño 30 de una población de 3.000 elementos. En consecuencia, a los estratos que tienen un gran número de elementos se les debe asignar una muestra de tamaño grande.

La variabilidad es necesario considerarla porque para estimar un parámetro poblacional se requiere una muestra grande cuando la información no es homogénea.

Si el costo de obtener una observación no es igual en los diferentes estratos, como en el muestreo se debe minimizar el costo, es necesario tomar muestras pequeñas en los estratos donde se presentan altos costos en la recolección de la información.

Vamos a considerar unas situaciones que involucran estos tres factores.

Situación uno

Tamaño y afijación de la muestra que minimiza el costo de recolección de la información, y que tiene en cuenta las dispersiones de cada estrato las cuales son diferentes y que además tiene en cuenta el número de elementos de cada estrato. En estas circunstancias, el tamaño de la muestra está dado por la siguiente formula:

$$n = \frac{\left(\sum \sigma_i * w_i \sqrt{c_i}\right) * \left(\sum \frac{\sigma_i * w_i}{\sqrt{c_i}}\right)}{\frac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \frac{\sum \sigma_i^2 * w_i}{N}}$$

y la afijación de la muestra en los diferentes estratos está dada por:

$$n_i = n * \frac{\frac{\sigma_i * w_i}{\sqrt{c_i}}}{\sum \frac{\sigma_i * w_i}{\sqrt{c_i}}}$$

donde:

σ_i : es la desviación estándar del estrato poblacional i–ésimo.
c_i : es el costo para obtener una observación del estrato i–ésimo.
w_i : es la proporción del estrato poblacional i–ésimo.

Esta afijación se conoce como afijación óptima para costos, ya que balancea la variabilidad de cada estrato con su tamaño y donde se tiene en cuenta el costo de la recolección de la información.

Desarrollemos estas ideas con un ejemplo.

Ejemplo 3. El auditor de una empresa de servicios de telecomunicaciones desea hacer una estimación del valor promedio y del valor total del costo mensual del servicio de teléfono de las 4.000 familias del estrato I, II y III de cierta ciudad.

En el siguiente cuadro se presenta la configuración de los tres estratos poblacionales:

Estratos	Tamaño del estrato N_i	Proporción del estrato w_i
I	2400	60%
II	1200	30%
III	400	10%
	$N = 4000$	

Para calcular el tamaño de la muestra es necesario conocer la desviación estándar de cada estrato, pero como son desconocidos generalmente, serán estimadas con una muestra piloto, cuyo resultado fue el siguiente:

Estratos	Promedio muestral $\bar{X}_h$	Desviación muestral S_h
I	6800	2500
II	18000	12000
III	52000	31000

El valor de la media estratificada es:

$$\overline{X}_{est} = \sum_{1}^{3} \overline{X_h} * w_h = 14.680 \ pesos$$

Se sabe que el costo de obtener una observación en cada estrato es:

$$C_I = 1.000 \ pesos$$
$$C_{II} = 2.000 \ pesos$$
$$C_{III} = 5.000 \ pesos$$

Esta diferencia de los costos se explica por la ubicación física de los estratos la cual da gastos diferentes en transporte.

Con los anteriores elementos, con un error del 5% y una confianza del 95% podemos calcular el tamaño de la muestra, así:

$$n = \frac{\left(\sum S_i * w_i * \sqrt{c_i}\right) * \left(\sum \frac{S_i * w_i}{\sqrt{c_i}}\right)}{\frac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \frac{\sum S_i^2 * w_i}{N}}$$

En el ejemplo, tenemos los siguientes valores:

$$\sum S_i * w_i * \sqrt{c_i} = 427.634{,}1615$$

$$\sum \frac{S_i * w_i}{\sqrt{c_i}} = 171.77$$

$$E = 734$$

$$N = 4000$$

$$Z_{1-\frac{\alpha}{2}} = 1.96$$

$$\sum S_i^2 * w_i = 143.050.000$$

$$n = 418$$

La afijación óptima por costos para cada estrato es:

$n_i = n * \dfrac{\dfrac{S_i * w_i}{\sqrt{c_i}}}{\sum \dfrac{S_i * w_i}{\sqrt{c_i}}}$,	$n_1 = 115$ $n_2 = 196$ $n_3 = 107$

Según el anterior resultado, para estimar el valor promedio y el valor total del costo mensual en teléfono en las 4.000 familias, con un diseño muestral aleatorio simple, es necesario revisar el costo en teléfono de 115 familias del estrato I, 196 del estrato II y 107 del estrato III. La recolección de las facturas se hace de acuerdo al muestreo aleatorio simple o de acuerdo al muestreo sistemático. En esta situación para obtener este tamaño de muestra se tiene en cuenta: el número de familias de cada estrato, la variabilidad de cada estrato y el costo de recolectar la información en cada estrato.

Situación dos

Tamaño y afijación de la muestra cuando los costos para obtener las observaciones de cada estrato son iguales; por tanto se tiene en cuenta la dispersión de cada estrato los cuales son diferentes y el número de elementos de cada estrato.

En esta situación, el tamaño de la muestra está dado por:

$$n = \frac{\left(\sum \sigma_i * w_i\right)^2}{\dfrac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \dfrac{\sum \sigma_i^2 * w_i}{N}}$$

y la afijación de la muestra en los diferentes estratos está dada por:

$$n_i = n * \frac{\sigma_i * w_i}{\sum \sigma_i * w_i}$$

Esta afijación es directamente proporcional al producto del tamaño del estrato por su variabilidad y es llamada afijación óptima. Veamos esta situación en el siguiente ejemplo:

Ejemplo 4. El auditor de la empresa de telecomunicaciones del ejemplo anterior, decide utilizar el teléfono para recoger la información de la muestra piloto para reducir costos. Por tanto el costo de obtener una observación es el mismo en cada estrato. Con la misma información del ejemplo 3, el tamaño de la muestra es:

$$n = \frac{\left(\sum S_i * w_i\right)^2}{\dfrac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \dfrac{\sum S_i^2 * w_i}{N}}$$

como:

$$\left(\sum S_i * w_i\right)^2 = 24'820000$$

el tamaño de la muestra es: $n = 382$ y la afijación óptima es:

$n_i = n * \dfrac{S_i * w_i}{\sum S_i * w_i}$,	$n_1 = 70$ $n_2 = 168$ $n_3 = 144$

Según este resultado, teniendo en cuenta el número de familias de cada estrato y la variabilidad interior de cada estrato, para estimar el costo promedio y total mensual de las familias de estos tres estratos, el auditor debe revisar 70 facturas del estrato I, 168 facturas del estrato II y 144 facturas del estrato III. La selección de estas facturas en cada estrato se hace de acuerdo al muestreo aleatorio simple o de acuerdo al muestreo sistemático.

Situación tres

Tamaño y afijación de la muestra cuando los costos para obtener las observaciones de cada estrato son iguales y cuando las varianzas de los *diferentes estratos son iguales; por tanto se tiene en cuenta el número de elementos de cada estrato.* En estas circunstancias el tamaño de la muestra está dado por la siguiente formula:

$$n = \frac{\sum \sigma_i^2 * w_i}{\frac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \frac{\sum \sigma_i^2 * w_i}{N}}$$

y la afijación de la muestra en los diferentes estratos está dada por: $n_i = n * w_i$ la asignación de la muestra es directamente proporcional al tamaño de cada estrato y es llamada afijación proporcional.

Ejemplo 5. El auditor de la empresa de telecomunicaciones del ejemplo 3, decide recoger la información de la prueba piloto por teléfono, por lo cual el costo de obtener una observación de cada estrato es el mismo. Además, la desviación estándar de cada estrato, fue estimada en 6.000 de acuerdo a los resultados de una muestra preliminar revisada.

Con un error del 5% y una confianza del 95%, hallar el tamaño de la muestra.

Como los costos de recolectar la información en los diferentes estratos es igual y como las varianzas de los estratos son iguales, el tamaño de la muestra está dado por:

$$n = \frac{\sum S_i^2 * w_i}{\frac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \frac{\sum S_i^2 * w_i}{N}} = 813$$

y la afijación proporcional es:

$n_i = n * w_i$,	$n_1 = 488$ $n_2 = 244$ $n_3 = 81$

Según estos resultados, para estimar el valor promedio y total del costo mensual en teléfono es necesario recolectar 488 facturas del estrato I, 244 del estrato II y 81 del estrato III. La recolección de la muestra en cada estrato se hace de acuerdo al diseño muestral aleatorio simple o de acuerdo al diseño muestral sistemático.

2.2. Tamaño de la muestra para la estimación de la proporción poblacional cuando la variable relevante es cualitativa

Como lo hicimos en la determinación del tamaño de la muestra para la determinación de la media y del total poblacional consideraremos las tres situaciones que involucran los tres factores siguientes: El número de elementos en cada estrato. La variabilidad de las observaciones dentro de cada estrato. El costo por obtener una observación de cada estrato.

Situación uno

Tamaño y afijación de la muestra que minimiza el costo de recolección de la información y que tiene en cuenta las dispersiones de cada estrato las cuales son diferentes y que además tiene en cuenta el número de elementos de cada estrato.

En estas circunstancias el tamaño de la muestra está dado por:

$$n = \frac{\left(\sum \sqrt{p_i(1-p_i)} * w_i * \sqrt{c_i}\right) * \left(\sum \frac{\sqrt{p_i(1-p_i)} * w_i}{\sqrt{c_i}}\right)}{\frac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \frac{\sum p_i(1-p_i) * w_i}{N}}$$

donde $p_i(1-p_i)$: es la varianza del estrato i–ésimo.

Y la afijación de la muestra es:

$$n_i = n * \frac{\frac{\sqrt{p_i(1-p_i)} * w_i}{\sqrt{c_i}}}{\sum \frac{\sqrt{p_i(1-p_i)} * w_i}{\sqrt{c_i}}}$$

Esta afijación es llamada la afijación óptima para costos.

Veamos el siguiente ejemplo.

Ejemplo 6. Una empresa de servicios públicos entre los productos que ofrece a la comunidad está el servicio de Internet. El auditor de esta empresa quiere hallar el tamaño de la muestra para estimar el porcentaje de las 6.000 familias de los estratos IV, V y VI que poseen Internet.

La distribución de las familias en estos tres estratos es la siguiente:

Estratos	Tamaño del estrato N_i	Proporción del estrato w_i
IV	3.000	0.50
V	1.800	0.30
VI	1.200	0.20
	$N = 6.000$	

Para calcular el tamaño de la muestra es necesario conocer la desviación estándar de cada estrato, la cual es desconocida, pero fue estimada por el auditor mediante una muestra piloto, cuyo resultado fue el siguiente:

Estratos	$\hat{p}_i$	$1-\hat{p}_i$
IV	0.40	0.60
V	0.80	0.20
VI	0.90	0.10

Donde $p_i(1-p_i)$ es el estimador de la proporción poblacional del estrato i.

Además, se sabe que el costo de obtener una observación en cada estrato:

$$C_{IV} = 500$$
$$C_{V} = 600$$
$$C_{VI} = 1.000$$

Esta diferencia en costos está justificada por la geografía física de cada estrato, donde los costos de transporte son diferentes.

Con un error del 5% y una confianza del 95%, podemos calcular el tamaño de la muestra así:

$$n = \frac{\left(\sum \sqrt{p_i(1-p_i)} * w_i * \sqrt{c_i}\right) * \left(\sum \frac{\sqrt{p_i(1-p_i)} * w_i}{\sqrt{c_i}}\right)}{\frac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \frac{\sum p_i(1-p_i) * w_i}{N}}$$

En el ejemplo tenemos los siguientes valores:

$$\sum \sqrt{p_i(1-p_i)} * w_i * \sqrt{c_i} = 10.2973$$

$$E = 0.05$$

$$Z_{1-\frac{\alpha}{2}} = 1.96$$

$$N = 6.000$$

$$\sum \frac{\sqrt{p_i(1-p_i)} * w_i}{\sqrt{c_i}} = 0.018$$

$$\sum p_i(1-p_i) * w_i = 0.186$$

Luego, el tamaño de la muestra es $n = 269$.

Y la afijación óptima por costos para cada estrato es:

$n_i = n * \frac{\frac{\sqrt{p_i(1-p_i)} * w_i}{\sqrt{c_i}}}{\sum \frac{\sqrt{p_i(1-p_i)} * w_i}{\sqrt{c_i}}}$,	$n_1 = 166$ $n_2 = 75$ $n_3 = 30$

Un el anterior resultado, para estimar el porcentaje de familias que tienen el servicio de Internet, con un diseño muestral aleatorio simple, es necesario analizar 166 familias del estrato IV, 75 del estrato V y 30 del estrato VI. El auditor para seleccionar las familias, en cada estrato utiliza el muestreo aleatorio simple o el muestreo sistemático. Para obtener éste tamaño de muestra se tuvo en cuenta: el número de familias de cada estrato, la variabilidad de cada estrato y el costo de recolectar la información en cada estrato.

Situación dos

Tamaño y afijación de la muestra cuando los costos para obtener las observaciones de cada estrato son iguales; por tanto se tiene en cuenta la dispersión de cada estrato, las cuales son diferentes y el número de elementos de cada estrato.

En estas circunstancias, el tamaño de la muestra está dado por:

$$n = \frac{\left(\sum \sqrt{p_i(1-p_i)} * w_i\right)^2}{\dfrac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \dfrac{\sum p_i(1-p_i) * w_i}{N}}$$

Y la afijación de la muestra en los diferentes estratos está dada por:

$$n_i = n * \frac{\sqrt{p_i(1-p_i)} * w_i}{\sum \sqrt{p_i(1-p_i)} * w_i}$$

Esta afijación es directamente proporcional al producto del tamaño del estrato por su variabilidad y es llamada afijación óptima.

Ejemplo 7. El auditor de la empresa de servicios públicos del ejemplo anterior decide utilizar el teléfono para recoger la información de la muestra piloto a fin de reducir costos; en consecuencia, el costo de obtener una información es igual en cada estrato.

En este caso, el tamaño de la muestra es:

$$n = \frac{\left(\sum \sqrt{\hat{p}_i(1-\hat{p}_i)} * w_i\right)^2}{\frac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \frac{\sum \hat{p}_i(1-\hat{p}_i) * w_i}{N}}$$

Utilizando la información del ejemplo 6, tenemos los siguientes valores:

$$\sum \sqrt{\hat{p}_i(1-\hat{p}_i)} * w_i = 0.42495$$

$$\sum \hat{p}_i(1-\hat{p}_i) * w_i = 0.186$$

$$E = 0.05$$

$$Z_{1-\frac{\alpha}{2}} = 1.96$$

Luego, el tamaño de la muestra es $n = 267$ y la afijación óptima es:

$$n_i = n * \frac{\sqrt{\hat{p}_i(1-\hat{p}_i)} * w_i}{\sum \sqrt{\hat{p}_i(1-\hat{p}_i)} * w_i}, \qquad \begin{aligned} n_1 &= 154 \\ n_2 &= 76 \\ n_3 &= 37 \end{aligned}$$

Según este resultado, teniendo en cuenta el número de familias de cada estrato y la variabilidad interior de cada estrato, para estimar el porcentaje de familias que poseen el servicio de Internet es necesario analizar 154 facturas del estrato IV, 76 del estrato V y 37 del estrato VI. El auditor para seleccionar las familias de cada estrato, utiliza el muestreo aleatorio simple o el muestreo sistemático.

Situación tres

Tamaño y afijación de la muestra cuando los costos para obtener las observaciones de cada estrato son iguales y cuando las varianzas de los *diferentes estratos son iguales. Por tanto se tiene en cuenta el número de elementos de cada estrato.*

En estas circunstancias el tamaño de la muestra está dado por la siguiente formula:

$$n = \frac{\sum \sqrt{p_i(1-p_i)} * w_i}{\dfrac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \dfrac{\sum p_i(1-p_i) * w_i}{N}}$$

y la afijación de la muestra en los diferentes estratos está dada por:

$$n_i = n * w_i$$

La asignación de la muestra es directamente proporcional al tamaño de cada estrato y es llamada afijación proporcional. Veamos el siguiente ejemplo.

Ejemplo 8. El auditor de la empresa de servicios públicos decide recoger la información de la prueba inicial por teléfono, por lo cual el costo de obtener una observación de cada estrato es el mismo. Además, la varianza de cada estrato es igual, debido a que la proporción de familias que tienen el servicio de Internet fue estimado en 0.60 para los tres estratos, según los resultados de una prueba preliminar revisada. Con un error del 5% y una confianza del 95% hallar el tamaño de la muestra.

Como los costos de recolectar la información en los diferentes estratos es igual y como las varianzas de los estratos son iguales, el tamaño de la muestra se halla con la siguiente formula:

$$n = \frac{\sum \sqrt{\hat{p}_i(1-\hat{p}_i)} * w_i}{\dfrac{E^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \dfrac{\sum \hat{p}_i(1-\hat{p}_i) * w_i}{N}}$$

Utilizando los valores obtenidos anteriormente en la situación uno y dos, el tamaño de la muestra es: $n = 273$ y la afijación proporcional:

$n_i = n * w_i$,	$n_1 = 137$ $n_2 = 82$ $n_3 = 55$

Teniendo en cuenta el número de familias en cada estrato, para estimar el porcentaje de las 6.000 familias que poseen el servicio de Internet, el auditor debe analizar 137 familias del estrato IV, 82 familias del estrato V y 55 familias del estrato VI. La selección de las familias de cada estrato se hace con el muestreo aleatorio simple o con el muestreo sistemático.

3. Tamaño de la muestra en el diseño muestral aleatorio sistemático

Si el orden de la información es aleatorio, las formulas para hallar el tamaño de la muestra son las mismas del muestreo aleatorio simple. Sin embargo, en muchas situaciones el supuesto de que el orden de la información es aleatorio puede ser falso lo que nos llevaría a calcular un tamaño de muestra totalmente equivocado. Para evitar éste inconveniente se trabaja con un muestreo sistemático replicado, en el cual no es necesario hacer ningún supuesto sobre el orden de la información.

4. Tamaño de la muestra en el diseño muestral sistemático replicado

4.1. Tamaño de la muestra para la estimación de la media y del total poblacional cuando la variable relevante es cuantitativa

El procedimiento consiste en tomar una muestra piloto de muestras sistemáticas replicadas del marco muestral, con la cual se estiman los parámetros poblacionales y así determinar el número de muestras necesarias replicadas que garanticen la confiabilidad y precisión requeridas. La formula para hallar el tamaño de la muestra es:

$$n = \frac{\dfrac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 * Var(\overline{Y}_i)}{\left(\overline{Y}\right)^2} * \dfrac{N}{n}}{E^2 * \left(\dfrac{N}{n} - 1\right) + \left(Z_{1-\frac{\alpha}{2}}\right)^2 * \dfrac{Var(\overline{Y}_i)}{\overline{Y}^2}}$$

Donde: $\overline{Y_i}$: es la media de la muestra sistemática i–ésima.

$$\overline{Y_i} = \sum_{1}^{n_s} \frac{\overline{Y_i}}{n_s'}$$

$$Var(\overline{Y}_i) = \frac{\sum (\overline{Y}_i - \overline{Y})^2}{n_s' - 1}$$

Veamos un ejemplo.

Ejemplo 9. Una asociación de pequeños comerciantes tiene 500 afiliados. El auditor de esta asociación quiere estimar el promedio y el total de las ventas diarias de 500 almacenes; para este fin decide aplicar el muestreo sistemático replicado y para poder calcular el tamaño de la muestra toma una muestra preliminar piloto de 5 muestras sistemáticas replicadas de 5 elementos cada una. Para el cálculo del tamaño muestral el auditor decide un error del 5% y una confianza del 95%.

Solución 3. El valor de k es: $k = \frac{500}{5} = 100$

Para determinar los puntos de inicio aleatorio para las 5 muestras sistemáticas se generan 5 números aleatorios entre 1 y 100 cuyos resultados fueron: 74, 19, 83, 7, 38.

Las 5 muestras sistemáticas replicadas preliminares fueron:

Muestra uno:

Orden de las facturas de venta	7	107	207	307	407
Valores de la factura	284.000	414.000	394.000	358.000	360.000

Muestra dos:

Orden de las facturas de venta	19	119	219	319	419
Valores de la factura	351.000	342.000	462.000	371.000	382.000

Muestra tres:

Orden de las facturas de venta	38	138	238	338	438
Valores de la factura	479.000	355.000	262.000	442.000	385.000

Muestra cuatro:

Orden de las facturas de venta	74	174	274	374	474
Valores de la factura	181.000	416.000	336.000	306.000	310.000

Muestra cinco:

Orden de las facturas de venta	83	183	283	383	483
Valores de la factura	193.000	293.000	374.000	493.000	393.000

Las medias muestrales son:

$$\overline{Y}_1 = 362.000$$
$$\overline{Y}_2 = 381.000$$
$$\overline{Y}_3 = 384.600$$
$$\overline{Y}_4 = 309.800$$
$$\overline{Y}_5 = 349.200$$

Además, en esta muestra preliminar tenemos los siguientes valores:

$$Var\left(\overline{Y}_i\right) = 30.211,5$$
$$\overline{Y} = 357.320$$

Remplazando los valores anteriores en la formula del tamaño muestral y teniendo en cuenta que el error se toma en porcentaje, tenemos:

$$n = \frac{\dfrac{(1.96)^2 * (30.211,5)^2}{(357.320)^2} * \left(\dfrac{500}{5}\right)}{0.05^2 * \left(\dfrac{500}{5} - 1\right) + (1.96)^2 * \dfrac{(30.211,5)^2}{(357.320)^2}} = 10$$

Por lo tanto, el auditor para estimar el valor promedio y total de las ventas diarias de los 500 almacenes, debe replicar 5 muestras sistemáticas adicionales.

5. Tamaño de la muestra en el diseño muestral por conglomerados

5.1. Tamaño de la muestra para la estimación de la media y del total poblacional cuando la variable relevante es cuantitativa

Para estimar la media poblacional con un límite de error E y con una confianza de $(1-\alpha)\%$, el número de conglomerados que es necesario muestrear está dado por:

$$n = \frac{N * \sigma_c^2}{N * \dfrac{E^2 * \overline{M}^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \sigma_c^2}$$

Donde:

N : número de conglomerados de la población.

$\overline{M}$: tamaño promedio del conglomerado poblacional, el cual como frecuentemente es desconocido se estima por $\overline{M} = \dfrac{\sum m_i}{n}$.

$\sigma_c^2 = \dfrac{\sum (Y_i - \overline{Y} * M_i)^2}{N-1}$, el cual generalmente es desconocido y es necesario estimarlo en base a estudios anteriores o de acuerdo a los resultados de una muestra piloto por $S_c^2 = \dfrac{\sum\left(Y_i - \overline{Y} * m_i\right)^2}{n-1}$.

Veamos esta situación en el siguiente ejemplo.

Ejemplo 10. Un auditor va a estimar el promedio de los gastos diarios de teléfono por familia en cierto barrio de una ciudad y además va a estimar el porcentaje de estas familias que poseen servicio de Internet. Para ello decide utilizar un diseño muestral por conglomerados. El barrio consta de 80 manzanas y generando 12 números aleatorios del 1 al 80 selecciona las 12 manzanas que constituyen la muestra aleatoria preliminar.

En el siguiente cuadro aparece el resultado muestral de la muestra piloto, donde:

m_i : número de residencias de la manzana i.

Y_i : el valor total de los gastos diarios en teléfono en el conglomerado i.

a_i : número de familias que tienen Internet en el conglomerado i.

Conglomerado i	m_i	a_i	Y_i
1	25	15	1.250
2	20	12	1.000
3	18/	10	720
4	24	14	960
5	25	13	1.500
6	20	123	1.200
7	15	10	750
8	18	10	720
9	20	11	1.200
10	25	15	1.300
11	22	16	1.400
12	24	15	1.3501

El auditor quiere calcular el número de conglomerados que debe muestrear para estimar el valor promedio y total de las gastos diarios en teléfono de las 2.000 familias del barrio de interés. La estimación se hace con un error del 5% y una confianza del 95%.

Solución 4. De la información de la muestra preliminar tenemos los siguientes valores:

$$N = 80$$
$$n = 12$$
$$\bar{m} = \frac{\sum m_i}{n} = 21.33$$
$$E = 3$$
$$S_c^2 = 31621{,}80$$

El error es aproximadamente un 5% del gasto diario promedio de teléfonos por familia. El número de conglomerados a muestrear está dado por:

$$n = \frac{N * S_c^2}{N * \dfrac{E^2 * \bar{m}^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + S_c^2} = 22$$

Por tanto, el auditor para estimar el valor promedio y total de los gastos diarios en teléfono de las 2.000 familias debe muestrear 10 conglomerados adicionales.

9.2. Tamaño de la muestra para la estimación de la proporción poblacional cuando la variable relevante es cualitativa

El tamaño de la muestra para estimar la proporción poblacional, p, con un limite de error E y una confianza del $(1-\alpha)\%$ está dado por:

$$n = \frac{N * \sigma_c^2}{N * \dfrac{E^2 * \overline{M}^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2} + \sigma_c^2}$$

donde σ_c^2 se estima por:

$$S_c^2 = \frac{\sum (a_i - \hat{p} * m_i)^2}{n-1}$$

Veamos el siguiente ejemplo:

Ejemplo 11. De acuerdo a la muestra preliminar piloto del ejemplo anterior, hallar el número de conglomerados que tiene que incluir el auditor en la muestra, para estimar, con un 5% de error y una confianza del 95%, el porcentaje de las 2.000 familias que tienen Internet.

Solución 5. De los resultados de la muestra preliminar tenemos los siguientes valores:

$$N = 80$$
$$n = 12$$
$$\overline{m} = 21.33$$
$$S_c^2 = 1.4171$$

De modo que el tamaño de la muestra es:

$$n = \frac{80 * 1.4171}{\dfrac{80 * (0.05)^2 * (21.33)^2}{(1.96)^2} + 1.4171} = 5$$

Como el tamaño muestral es muy bajo el auditor decide aumentar el tamaño de la muestra y para ello decide reducir el error de estimación , así: con un error del 3%, el tamaño de la muestra es 12 conglomerados. En este caso la muestra preliminar sería la muestra definitiva y si decide el auditor un error del 2%, el tamaño de la muestra sería de 22 conglomerados, por tanto el auditor para estimar el porcentaje de familias que poseen Internet tendría necesidad de muestrear 10 conglomerados adicionales.

Con el estudio de la muestra en los diferentes diseños muestrales, hemos dado respuesta a la segunda gran pregunta del muestreo: *cuál debe ser el tamaño de la muestra aleatoria.*

Bibliografía

Nota del editor: las fuentes bibliográficas pueden consultarse en la Facultad de Ciencias Económicas de la Universidad de Antioquia, en el libro del mismo autor "Estadísticas para las ciencias contables", en proceso de publicación.