UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

APUNTES Y PROPUESTA DE EJERCICIOS PARA LA CLASE DE MUESTREO

T E S I S

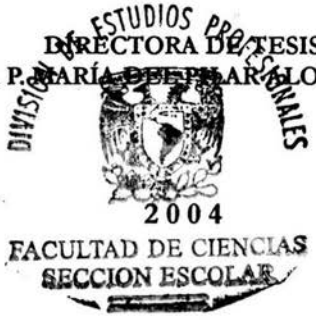
QUE PARA OBTENER EL TÍTULO DE:

A C T U A R Í A

P R E S E N T A :

LAURA FUENTES SÁNCHEZ

FACULTAD DE CIENCIAS
UNAM



DIRECTORA DE TESIS:
M. en A. P. MARÍA DEL PILAR ALONSO REYES

2004

UNIVERSIDAD NACIONAL
AVENIDA DE
MEXICO

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

"Apuntes y propuesta de ejercicios para la clase de muestreo"

realizado por Fuentes Sánchez Laura

con número de cuenta 09550297-3 , quien cubrió los créditos de la carrera de: Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis
Propietario M. en A.P. María del Pilar Alonso Reyes

Propietario M. en C. José Antonio Flores Díaz

Propietario Dr. Luis Antonio Rincón Solís

Suplente Act. Lucio Gerardo Chávez Heredia

Suplente Act. Jaime Vázquez Alamilla

Consejo Departamental de
Matemáticas

M. en C. José Antonio Flores Díaz

FACULTAD DE CIENCIAS
CONSEJO DEPARTAMENTAL
DE
MATEMATICAS

DEDICATORIAS.

A mis padres
Por todo el apoyo, atención y cuidado que siempre me han brindado ya que fueron un soporte y ejemplo para lograr esta meta, ya que sin ellos no habría sido posible alcanzarla en mi vida.

A mi esposo
Por el apoyo, cariño, amor y compañía que me ha brindado a lo largo de nuestra relación ya que siempre me ha motivado a seguirme superando día tras día.

A mis hijos
A pesar de que son aún pequeños han sido un gran motivo de superación personal y profesional.

A mis hermanos
Con los cuales he pasado momentos felices y que sea esta, un punto de referencia para que continuamente se sigan superando día con día y alcancen sus metas que se hayan propuesto en la vida.

A mis abuelos
Por su apoyo, amor y comprensión y por ayudar a mantener la familia unida.

A mi directora de tesis María del Pilar Alonso Reyes
Por su orientación, paciencia, tiempo, dedicación y observaciones al presente trabajo.

A mis sinodales
Gracias por aceptar su participación y comentarios en le elaboración de esta tesis.

Laura Fuentes Sánchez.

ÍNDICE.

Capítulo 4 : Muestreo por conglomerados.

4.1 Muestreo por conglomerados.

4.2 Selección aleatoria de conglomerados.

4.3 Muestreo conglomerado de una etapa: conglomerados de tamaños desiguales.

4.4 Selección con probabilidades desiguales y con restitución.

4.5 Ejercicios propuestos.

CAPÍTULO 1:

INTRODUCCIÓN.

1.1 INTRODUCCIÓN.

La intención al hacer esta tesis es dar a los alumnos de muestreo, y a todas las personas interesadas en este tema, una referencia en la cual se puedan apoyar para revisar algunos aspectos del muestreo ya que el tema tiene una gran importancia. Primeramente se hablará de manera resumida de lo que es el muestreo aleatorio simple, muestreo aleatorio estratificado y en donde se pondrá especial atención es en el muestreo por conglomerados.

El muestreo desempeña un papel de vital importancia en el diseño de investigación de poblaciones, las cuales pueden ser animales, plantas, personas, cuentas bancarias, en fin un sin número de tipos de poblaciones. La importancia se debe a que si la población es muy grande, genera el consumo de mucho tiempo, dinero y esfuerzo, así una muestra implica resultados más rápidos para la variable o variables en cuestión, teniendo una ganancia significativa en los factores antes mencionados, pero estas ganancias implicarían pérdidas en la precisión de los estimadores.

El **muestreo aleatorio** es un método para escoger a un conjunto de elementos de una población sobre la cual se va ha llevar a cabo una investigación; implica seleccionar los sujetos de manera que cada elemento en la población tenga la misma oportunidad de ser escogida para el estudio. Formalmente, se puede decir que el muestreo aleatorio simple es un método de selección de n unidades sacadas de N, de tal manera que cada una de las posibles muestras de tamaño n, tienen la misma oportunidad de ser escogidas, donde ${}_NC_n$, es una fórmula combinatoria que indica que si todas las unidades de la población pueden ser distinguidas unas de otras, el número de muestras distintas de tamaño n que pueden ser sacadas de N unidades es ${}_NC_n$.

La **muestra** es la porción de una población estudiada en la investigación y una **muestra representativa** es una muestra que refleja las características importantes de la población que se estudia.

La **población** es una colección o agrupación de objetos o de entes que como ya se mencionó anteriormente pueden ser animales, plantas, personas, cuentas bancarias, etc. que se caracterizan por poseer ciertas propiedades específicas. La **población objetivo** es el conjunto de elementos bajo estudio. La **población a estudiar** es aquella sobre la que se desea efectuar inferencias y queda definida antes de iniciar el trabajo de campo. La **población muestreada** es la agrupación de donde se sacan los resultados y conclusiones que se hacen sobre la muestra.

La **unidad de muestreo** es el elemento que participa en el proceso de selección de la muestra con una probabilidad conocida.

Un **marco de muestreo** o **marco poblacional** es la relación de elementos que pertenecen a la población objetivo, éste debe de poseer datos que permitan la identificación y localización de los elementos relacionados.

Para llevar a cabo un muestreo probabilístico, es necesario contar con este marco, en ausencia de éste, surge el muestreo no probabilístico, el cual puede ser:

a) *Muestreo por cuotas*, es el muestreo que establece cuotas. Se entenderá por cuota un conjunto de características especificadas en la población.
b) *Muestreo de juicio o de selección intencional*, en éste se muestrea bajo la subjetividad del encuestador, generando elementos muestrales a criterio del encuestador.

En el muestreo probabilístico, se puede definir el conjunto de muestras $\delta_1, \delta_2, \ldots, \delta_v$ que el procedimiento es capaz de elegir si se aplica a una población específica, lo cual indica que se pueden decir con precisión cuales son las unidades de muestreo que pertenecen a $\delta_1, \delta_2, \ldots, \delta_v$.

Cada muestra probable δ_i tiene asignada una probabilidad de selección Π_i

Se selecciona una de las δ_i, por un proceso aleatorio en el que cada δ_i tiene probabilidad Π_i de ser elegida.

El método para calcular la estimación a partir de la muestra debe ser definido y debe conducir a una estimación única para cualquier muestra específica.

Este tipo de muestreo es muy importante ya que su medibilidad, la cual permite calcular estimaciones válidas a partir de la muestra, lleva a inferencias estadísticas objetivas; además, permite mejoramientos acumulativos a través de la separación y la evaluación objetiva de sus fuentes de errores.

1.2 MÉTODOS DE INVESTIGACIÓN

Los métodos de investigación permiten describir, detallar y analizar un problema a través de técnicas estadísticas (como la regresión, el muestreo, etc.).

Lo primero que se tiene que hacer es definir el problema y decir cual es el objetivo o interés particular, por lo tanto, se tienen que definir los objetivos y plantear las hipótesis.

La forma de obtener o recopilar la información que interesa de las muestras, de manera rápida y económica, son los cuestionarios, los cuales proporcionan información accesible y rápida y las respuestas no requieren pensarse mucho. Para responder, los participantes (encuestados) simplemente señalan la respuesta adecuada.

Un cuestionario consta de varias partes:

-**Carátula**: Identificación del cuestionario, esto es, quien lo está aplicando, introducción, e identificación de la persona o folio.

-**Bloque de preguntas**: Existen diferentes tipos de preguntas; las **preguntas abiertas** son en las que el entrevistado puede explayarse todo lo que quiera, aquí el problema es que no se le puede aplicar un sistema de codificación tan fácilmente; las **preguntas cerradas o estructuradas**, se les conoce también como de opción múltiple, son en las que se mencionan al entrevistado varias respuestas, y éste elige una. Estas preguntas pueden ser: dicotómicas cuando se tienen dos opciones o en abanico cuando se tienen más de dos opciones; las **preguntas combinadas**, son en las que la primera parte es una pregunta cerrada y posteriormente el entrevistado tiene que dar su punto de vista.

El éxito de los cuestionarios depende de una serie de factores, como son:

- No hay que omitir preguntas obvias.
- La forma en que se redactan las preguntas debe ser simple y específica para que el significado sea claro, y van de las preguntas más sencillas a las más complicadas, el lenguaje tiene que ser un lenguaje coloquial.
- Hay que definir exactamente todos los términos de las preguntas para que esto no se preste a confusiones.
- Las opciones para cada pregunta deben reflejar adecuadamente la amplitud de las posibles respuestas. Las preguntas no deben implicar preferencias hacia o en contra de opciones de respuesta específicas.
- Las preguntas deben redactarse de manera que eviten responderlas sin poner atención, y hay que evitar las preguntas que deban ser contestadas por la memoria.

El cuestionario tiene una parte en donde el entrevistado contesta las preguntas ya sea con una "x", con un número, con una letra o con sus propias palabras dependiendo del tipo de pregunta que se tenga y otra parte que generalmente se pone sombreada para que no se conteste en esa área, que es en donde el entrevistador realiza las codificaciones después de que se haya contestado el cuestionario.

Una forma de aplicar un cuestionario es que los entrevistadores vayan y entrevisten a las personas seleccionadas, pero nadie asegura que el entrevistador no falsificó los datos o que él sólo respondió los cuestionarios, además de que el costo es alto debido a que se tienen que pagar viáticos si el entrevistador tiene que trasladarse a otra ciudad. Otra forma de aplicar un cuestionario es enviarlo por correo, pero aquí la desventaja que se tiene es que no todas las personas van a contestar o a enviar de regreso el cuestionario contestado, además no se sabe si la persona lo contestó con o sin influencia de otra persona, pero la ventaja de aplicarlo de esta forma es que el costo es menor, pero por otro lado se tiene la no

representatividad ya que no se escoge a las personas que van a regresar el cuestionario contestado.

Los tipos de diseño de encuesta de acuerdo con la evolución del fenómeno estudiado son:

1.- **DISEÑO TRANSVERSAL** en el cual se miden en una sola ocasión la o las variables. No se considera la evolución temporal de las unidades o elementos estudiados, estas encuestas resultan generalizables a la población muestreada, pero sólo en el momento temporal en que ha tenido lugar el estudio.
La ventaja de aplicar un diseño transversal es que el costo es menor y el tiempo requerido es poco; hay una mayor facilidad para la obtención de datos y la posibilidad de utilizarlos aun cuando sólo se puede observar al sujeto una sola vez. El proceso de operaciones y codificación de tablas es más rápido. El costo de este diseño es barato, y requiere de menor atención en la etapa de planeación.

Las desventajas de aplicar un diseño de este tipo es que si dan la información falsa, no se puede corregir, ya que la observación se realiza en un momento en el tiempo.

2.- **DISEÑO LONGITUDINAL** es el que se realiza en varias ocasiones. Implica el seguimiento para comparar la evolución en el tiempo de los elementos.
La ventaja de aplicar este diseño es que la cantidad de información es completa, ya que se estudia varias veces al mismo sujeto. Permite establecer con más exactitud los conceptos de la investigación, existe menos posibilidad de que la gente de información falsa, ya que ésta se recoge en el tiempo requerido, y las personas no hacen tanto uso de su memoria.

Algunas desventajas del diseño longitudinal es que es caro y que en muchas ocasiones los entrevistados, después de varias entrevistas se nieguen a contestar o en su defecto que el individuo fallezca o emigre.

Como se mencionó al principio del tema, las investigaciones, después de haber definido el problema y definir los objetivos, tienen suposiciones, las cuales se llamarán hipótesis, éstas siempre se presentan como enunciados, de la forma: *Si* . . . (variable *x*) *entonces* . . .(variable *y*) . Las hipótesis más simples dice que un evento, la variable *x*, influye, causa o contribuye a un segundo evento, la variable *y* . La hipótesis maneja las dos variables en formas diferentes. La variable *x*, se encuentra bajo el control del investigador, se puede manipular como se quiera (se le conoce como variable independiente). En cambio, la variable *y* depende de la primera, por lo cual se le conoce como variable dependiente.

Por otra parte, de nada sirve un trabajo de investigación si no se presentan conclusiones, por eso el objetivo de estos trabajos es poder sacar conclusiones o extrapolar los resultados de la muestra estudiada.

Como se acaba de mencionar, lo primero que se tiene que hacer para llevar a cabo una encuesta por muestreo es la definición de los objetivos e hipótesis, y con esto dar una idea de las conclusiones que se quieren obtener para así tener en cuenta que tipo de tabulaciones se realizarán, las cuales pueden ser tan complejas como el investigador quiera; también hay que definir la población de encuesta, para saber a quien o que se va a muestrear, aquí se define el marco poblacional, pero existen problemas como la no respuesta o que el marco esté incompleto ya sea que haya menos elementos en el marco o que haya más elementos (sobrecobertura), con esto se decide si se lleva a cabo el muestreo o no ya que si no se cuenta con el marco poblacional entonces no se puede saber a quien seleccionar para llevar a cabo el muestreo.

Pero para esto, se tiene que tener un proceso de selección, para obtener la muestra estudiada, en este proceso, se dan reglas mediante las cuales se incluyen o se excluyen los elementos de la población, la cual debe de estar definida en tiempo y espacio, por ejemplo decir que la población objetivo van a ser los estudiantes de la Facultad de Ciencias (espacio) de la generación 95 (tiempo).

Después de tener bien definido el marco poblacional y que el cuestionario está orientado a los objetivos de la investigación, se determina la unidad última de muestreo que es a "quién se le aplicará el cuestionario". Se tiene que definir la precisión con la cual se va a trabajar, ya que el tamaño de la muestra depende tanto de la confianza como de la precisión que se desee, por lo tanto se debe de encontrar una ecuación que relacione el tamaño de muestra n, con la precisión deseada, dicha ecuación, por lo general se encuentra en función de parámetros desconocidos de la población, los cuales se estiman. Cuando se desean presentar resultados en subdivisiones, se debe calcular por separado el tamaño de la muestra para cada subdivisión y tomar el tamaño de muestra total como la suma de los valores de los tamaños calculados para las subdivisiones. Para calcular el tamaño de muestra, se requiere tener conocimiento de S^2, la cual se puede obtener tanto de encuestas previas (que no es muy frecuente que se hayan realizado encuestas en poblaciones similares y con objetivos también similares) como de muestras pilotos, también no hay que olvidar calcular la no respuesta.

Si se realiza una encuesta piloto, se hace para verificar si el desarrollo de la encuesta y su diseño van por buen camino, o si hace falta algo que no se tomó en cuenta, como por ejemplo que se haya omitido una pregunta o que al marco poblacional falte delimitarlo más específicamente, o que simplemente no se haya puesto la suficiente atención al aspecto económico, ya que éste también es un factor importante en el diseño de muestreo.

Ya que todo esto está aprobado, se realiza lo que es el levantamiento de información y posteriormente la codificación, la cual se busca en tablas que previamente se hicieron, y se realiza la captura de los datos, nótese que en esta parte puede existir un tipo de error que no es error de muestreo, y que no se toma en cuenta, como lo es la mala captura de los datos o la codificación errónea.

Finalmente, se realiza un análisis completo en el cuál se puede decir las características y el comportamiento de las variables estudiadas y así hacer las inferencias y extrapolaciones deseadas de la población o poblaciones, ya que el estudio puede ser comparativo o descriptivo, entendiendo por comparativo estudiar a dos poblaciones con características distintas y hacer una comparación de las variables en cuestión además de contrastar hipótesis causales.

1.3 MUESTREO ALEATORIO SIMPLE.

El muestreo aleatorio simple (M.A.S.) es el método de selección de n unidades en un conjunto de N, el cual es el tamaño de la población, de tal modo que cada una de las posibles muestras distintas (las cuales son las combinaciones sin reemplazo de N elementos tomados de n en n) tengan la misma oportunidad de ser elegidas.

Como en las extracciones subsecuentes se descarta el extraído anteriormente, este método es sin restitución y la probabilidad de que se extraigan las n unidades especificadas es:

$$\frac{n}{N} * \frac{(n-1)}{(N-1)} * \frac{(n-2)}{(N-2)} * \cdots\cdots * \frac{1}{(N-n+1)} = \frac{n!(N-n)!}{(N)!} = \frac{1}{{}_N C_n}$$

Este método también se llama muestreo aleatorio sin restitución. Nótese que como el muestreo es sin reemplazo, el tamaño de la muestra n siempre va a ser menor que el tamaño de la población N. La muestra no son las primeras n unidades, sino que se seleccionan por un proceso aleatorio.

El interés por muestrear es medir una característica, a la cual se le va a llamar variable (y_i).

Las letras mayúsculas van a representar los valores de la población, y las minúsculas van a representar los de la muestra.

Los valores de la población son:

N = Tamaño poblacional.

$Y_1, Y_2, \ldots, Y_N$ = las características o variables de la población.

$Y = \sum_{i=1}^{N} Y_i = Y_1 + Y_2 + Y_3 + \cdots\cdots + Y_N$ = Total de la población para la variable Y.

$$\overline{Y} = \frac{Y_1 + Y_2 + \ldots\ldots + Y_N}{N} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{Y}{N}$$ = Media poblacional.

$$\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$$ = Varianza de los elementos de la población.

$$S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \frac{N}{N-1}\sigma_Y^2$$ = Varianza de los elementos de la población .

La diferencia de poner en el denominador *N-1* en lugar de *N* es debido a que los resultados toman una forma un poco más sencilla.

Los valores de la muestra son:

n = tamaño muestral.

$y_1, y_2, y_3, \ldots\ldots, y_n$ = las variables de la muestra.

$$y = \sum_{i=1}^{n} y_i = y_1 + y_2 + \ldots\ldots + y_n$$ = Total de la muestra para la variable y.

$$\overline{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y}{n}$$ = Media muestral.

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$ = Varianza muestral.

Con el símbolo ^ se estará diciendo que es una estimación muestral de una característica de la población.

Algunos de los estimadores que se usan son:

$\hat{\bar{Y}} = \bar{y}$ La media de la población estimada es igual a la media muestral.

$\hat{Y} = N\bar{y} = N\frac{\sum_{i=1}^{n} y_i}{n}$ Estimación del total de la población.

Aquí se puede observar el factor $\frac{N}{n}$, el cual indica cuantas veces cabe la muestra en la población y se le conoce como factor de expansión.

El inverso de este cociente es $f = \frac{n}{N}$ al cual se le conoce como la fracción de muestreo o fracción de corrección por finitud, es la razón del tamaño de la muestra respecto a la población. Esta fracción es por lo general muy pequeña, ya que la muestra n es muy pequeña comparada con el tamaño de la población total N y cuando la población total es infinita entonces la fracción de corrección por finitud no se toma en cuenta para hacer los cálculos, ya que el cociente $\frac{n}{N}$ tiende a cero cuando N tiende a infinito, por lo tanto siempre se considerará que el tamaño de la población es muy grande con respecto al tamaño de la muestra, y lo que se estaría haciendo es sobrestimar el error estándar de la estimación $\bar{y}$.
Existe una corrección para poblaciones finitas la cual es el factor *1-f*.

Un método de estimación es consistente cuando la estimación se vuelve exactamente igual al valor de la población cuando $n = N$, y lo más deseable de las estimaciones es que fueran consistentes, pero una estimación no consistente no necesariamente está mal sino que los resultados sólo se toman para cuando el tamaño muestral n es pequeño comparado con el tamaño de la población.

Una estimación es insesgada si el valor promedio de la estimación es exactamente igual al valor verdadero de la población, esto es, si el promedio de los promedios es el verdadero.

Aquí en este caso, no se sabe como se distribuye la población total, esto es, no se sabe si su distribución es normal, t-student, etc.

A continuación se mencionarán algunos teoremas que se utilizan con frecuencia:

- La media muestral $\bar{y}$ es un estimador insesgado de $\bar{Y}$. Esto es que $E(\bar{y}) = \bar{Y}$

  De aquí se puede decir que $\hat{Y} = N\bar{y}$ es un estimador insesgado del total de la población (Y).[1]

- La varianza de la media $\bar{y}$ de una muestra aleatoria simple, (la cual se denotará como m.a.s.) es:

  $Var(\bar{y}) = E[(\bar{y} - \bar{Y})^2] = \frac{S^2(N-n)}{nN} = \left(\frac{1-f}{n}\right)S^2$, donde f es la fracción de muestreo.[2]

A continuación se define la covarianza de dos medias ya que si se desea sacar la varianza de la suma de dos medias, ésta va a ser la suma de las varianzas de cada media menos dos veces la covarianza.

- Si y_i , x_i son dos variables definidas en toda unidad de la población, y $\bar{y}, \bar{x}$ son las medias derivadas de una muestra aleatoria simple de tamaño n, entonces su covarianza es:

  $Cov(\bar{x}, \bar{y}) = E[(\bar{y} - \bar{Y})(\bar{x} - \bar{X})] = \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})$[3]

---

[1] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[2] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[3] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

- Se tiene que el error estándar (desviación estándar) de $\bar{y}$ es:

$$ee_{\bar{y}} = \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N}} = \frac{S}{\sqrt{n}}\sqrt{1-f} \quad .$$

Y para el total de la población estimado ($\hat{Y}$), se tiene que el error estándar es:

$$EE_{\hat{Y}} = \frac{NS}{\sqrt{n}}\sqrt{\frac{(N-n)}{N}} = \frac{NS}{\sqrt{n}}\sqrt{1-f}\,.^{4}$$

Para comparar la precisión obtenida por el MAS con otros métodos de muestreo, para estimar el tamaño de la muestra que se necesita en una encuesta, y para estimar la precisión realmente obtenida en una encuesta que se haya terminado, se utilizan las fórmulas anteriores para los errores estándares de la estimación de la media y total de la población .

Como la varianza de la población S^2 no se conoce en la práctica, se estima con los datos de la muestra, por lo tanto:

- En una m.a.s., $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ es una estimación insesgada de $S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1}$.[5]

- La varianza de $\hat{Y} = N\bar{y}$, como un estimador del total de la población Y es:

$$Var(\hat{Y}) = E(\hat{Y} - Y)^2 = \frac{N^2S^2}{n}\frac{(N-n)}{N} = \frac{N^2S^2}{n}(1-f).^{6}$$

Esta varianza sirve para ver que tan bueno es el muestreo, ya que es una medida de dispersión del estimador $\bar{y}$, también sirve para comparar métodos de

[4] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[5] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[6] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

muestreo, ya que entre menor sea la varianza, el muestreo va a ser mejor. Esta varianza sirve para obtener el tamaño de la muestra n, tomando en cuenta que la población es muy grande y habiendo quitado el factor de corrección por finitud.

- De lo anterior se puede decir que las estimaciones insesgadas de las varianzas de $\bar{y}$ y $\hat{Y} = N\bar{y}$ son:

$$\hat{Var}(\bar{y}) = v(\bar{y}) = s_{\bar{y}}^2 = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{s^2}{n}(1-f)$$

[7]

$$\hat{Var}(\hat{Y}) = v(\hat{Y}) = s_{\hat{Y}}^2 = \frac{N^2 s^2}{n}\left(\frac{N-n}{N}\right) = \frac{N^2 s^2}{n}(1-f)$$

[7] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

1.4 ESTIMACIÓN POR INTERVALOS

Ahora bien, como las estimaciones por intervalos son mejores, en cuanto a su interpretación, que las puntuales debido a que es mejor decir, por ejemplo, que la media se encuentra en un intervalo que decir que es exclusivamente un punto dado, es necesario las expresiones de los intervalos de confianza de los parámetros.
Cuando la población es pequeña se supone que tiene una distribución t de Student, por otro lado, si la población es grande, lo cual generalmente es así, entonces se supone que la población se distribuye de forma normal con parámetros (0,1), con esto, al sacar los limites superiores ($\hat{\bar{Y}}_U$), e inferiores ($\hat{\bar{Y}}_L$) de la media son:

$$\hat{\bar{Y}}_U = \bar{y} + \frac{ts}{\sqrt{n}}\sqrt{1-f}$$

$$\hat{\bar{Y}}_L = \bar{y} - \frac{ts}{\sqrt{n}}\sqrt{1-f}$$

esto es que $\hat{\bar{Y}}$ está en el intervalo $\left[\bar{y} - \frac{ts}{\sqrt{n}}\sqrt{1-f}\right]$ a $\left[\bar{y} + \frac{ts}{\sqrt{n}}\sqrt{1-f}\right]$, esta longitud del intervalo depende del error estándar (s), del nivel de probabilidad que se desee tener (t) y del tamaño de la muestra (n), esta aproximación mejora conforme aumenta el tamaño de muestra y conforme la confianza es mayor los límites son más estrechos.

Y los límites superiores e inferiores del total de la población son:

$$\hat{Y}_U = N\bar{y} + \frac{tNs}{\sqrt{n}}\sqrt{1-f}$$

$$\hat{Y}_L = N\bar{y} - \frac{tNs}{\sqrt{n}}\sqrt{1-f}$$

donde t es el cuantil $1 - \frac{\alpha}{2}$, de una distribución t de Student (si el tamaño de la muestra n es menor que 50), con $n - 1$ grados de libertad, estos son los grados de libertad en la varianza estimada s^2; y si el tamaño de la muestra n es mayor a 50, entonces el cuantil se toma de una distribución normal (0,1).

Esto le da al investigador la probabilidad de confianza que se tiene, por ejemplo, si se escoge a $1 - \frac{\alpha}{2} = 0.95$ se podría decir con seguridad que de cada 100 que se escojan, 95 de estos caerán en el intervalo obtenido para la media.

1.5 MUESTREO PARA PROPORCIONES

Debido a que los resultados que se obtienen en las encuestas, generalmente indican la proporción de la población que tiene cierta característica, ya que los resultados que se dan al aplicar una entrevista son en términos de porcentajes, entonces lo que se hace es clasificar a la población, la cual se encuentra dentro de una clase C y también la que no se encuentra en esta clase.

Supóngase que a_i es una variable aleatoria que toma el valor de 1 si la i-ésima unidad está en la clase C y el valor de 0 si no lo está, dicho de otra manera, a_i es el número de veces que el i-ésimo valor cae en la muestra.

Sea A = número de elementos en la población que pertenecen a la clase C.

N = total de la población.

a = número de elementos en la muestra que pertenecen a la clase C.

n = tamaño de muestra.

$P = \frac{A}{N}$ = proporción de elementos de la población que pertenecen a la clase C.

$p = \frac{a}{n}$ = proporción muestral de elementos que pertenecen a la clase C.

Entonces se tendría que la media de la muestra es :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{N} a_i y_i$$

pero como a_i tiene el valor de 1 si está en la muestra y 0 si no lo ésta, entonces la media muestral es:

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ &= \frac{a}{n} \\ &= p\end{aligned}$$

Y para la media poblacional, de igual manera se tendría:

$$\begin{aligned}\bar{Y} &= \frac{1}{N}\sum_{i=1}^{N} y_i \\ &= \frac{A}{N} \\ &= P\end{aligned}$$

Nótese que la probabilidad de que $a_i = 1$ es $\frac{n}{N}$ y de que $a_i = 0$ es $1-\frac{n}{N}$, por lo tanto a_i es un variable binomial con parámetros (n,P) con $P = \frac{1}{N}$, de donde:

$$\begin{aligned}E(a_i) &= nP \\ &= n\frac{1}{N} \qquad \text{(Esperanza de la variable aleatoria } a_i\text{).} \\ &= \frac{n}{N}\end{aligned}$$

$$\begin{aligned}Var(a_i) &= nPQ \\ &= nP(1-P) \\ &= n\left(\frac{1}{N}\right)\left(1-\frac{1}{N}\right) \qquad \text{(Varianza de } a_i\text{).} \\ &= \frac{n}{N}\left(1-\frac{1}{N}\right)\end{aligned}$$

Para conocer la covarianza entre a_i y a_j, se enunciará a continuación un teorema que será útil:

- *Considere n ensayos repetidos independiente con posibles resultados s_1, s_2, s_3,....,s_{k+1}. Sea $p_j = P(s_j)$ en un ensayo en particular y sea x_i = número de ensayos(n) resultantes en s_j, con $j=1,2,....,k+1$*

$$Sea\ z_{1\alpha} = \begin{cases} 1\ si\ el\ \alpha - ésimo\ ensayo\ resulta\ en\ s_j. \\ 0\ en\ otro\ caso. \end{cases}$$

$$Entonces \quad x_j = \sum_{\alpha=1}^{n} z_{j\alpha} \ y\ la\ covarianza\ entre\ x_i\ y\ x_j\ es:$$

$$Cov(x_i, x_j) = Cov\left(\sum_{\alpha=1}^{n} z_{j\alpha}, \sum_{\beta=1}^{n} z_{j\beta}\right) = -np_i p_j.^{8}$$

De esto se tiene que: $Cov(a_i, a_j) = -n\left(\frac{1}{N}\right)\left(\frac{1}{N}\right) = -\frac{n}{N^2}$

Algunos resultados importantes son:

- La proporción muestral $p = \frac{a}{n}$ es una estimación insesgada de la proporción en la población $P = \frac{A}{N}$.[9]

- La varianza de p es:

$$Var(p) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right) \quad \text{donde } Q = 1 \quad P.^{10}$$

---

[8] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[9] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[10] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

- La varianza de la estimación del número total de unidades en la clase C $(\hat{A} = Np)$ es:

$$Var(\hat{A}) = \frac{N^2 PQ}{n}\left(\frac{N-n}{N-1}\right).$$ [11]

- Una estimación insesgada de la varianza de p, derivada de una muestra, es:

$$\hat{V}ar(p) = v(p) = \frac{N-n}{n-1}\left(\frac{pq}{N}\right).$$ [12]

- Una estimación insesgada de la varianza de $\hat{A} = Np$ es:

$$\hat{V}ar(\hat{A}) = v(\hat{A}) = \frac{N(N-n)}{n-1}pq.$$ [13]

Se puede tener una fórmula más sencilla si se ignora el factor de corrección por finitud, y se sustituye $N - n$ por N y también $n - 1$ por n, así :

$$\hat{V}ar(\hat{A}) = v(\hat{A}) = \frac{N^2}{n}pq$$

[11] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[12] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[13] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

1.6 PRECISIÓN DE LOS INTERVALOS DE CONFIANZA.

Se refiere a la magnitud de las desviaciones respecto a la media obtenida por la aplicación repetida del procedimiento de muestreo, esto es, se repite varias veces el muestreo para tratar de que sea más preciso el estimador poblacional que se quiera, y con esto se verá cuanto se dispersó la media poblacional.

La **exactitud** es que tan lejos se está del verdadero valor poblacional, pero para saber que tan lejos o que tan cerca se está, se tiene que conocer el verdadero valor poblacional.

La **confianza** es que tanto por ciento se acerca al verdadero valor poblacional. Y para esto se manejan los intervalos de confianza, los cuales son:

$$p \pm t^{1-\alpha/2} \sqrt{\frac{pq(N-n)}{N(n-1)}} = p \pm \underbrace{t^{1-\alpha/2} \sqrt{1-f} \sqrt{\frac{pq}{n-1}}}_{\text{precisión}}$$

1.6.1 COEFICIENTE DE VARIACIÓN.

El coeficiente de variación (*C.V.*) es:

$C.V. = \frac{\sqrt{Var(\bar{y})}}{\bar{y}}$, pero en el caso en el que se desconozca S^2 , lo que se hace es estimar al C.V. con:

$$\hat{C}.V. = \frac{\sqrt{\hat{V}ar(\bar{y})}}{\bar{y}}$$

Para el caso continuo de medias se tiene:

$$\hat{C}.V. = \frac{\sqrt{\frac{1-f}{n} - s^2}}{\bar{y}} = \frac{s}{\bar{y}}\sqrt{\frac{1-f}{n}}$$

Para el caso de proporciones:

$$\hat{C}.V. = \frac{\sqrt{\frac{(N-n)pq}{N(n-1)}}}{p} = \sqrt{\frac{q}{p}}\sqrt{\frac{N-n}{N(n-1)}}$$

Esta expresión se utiliza mucho para sacar el tamaño de muestra, ya que se pide, por ejemplo, un coeficiente de variación de 0.5 y entonces se conoce todo menos n y con esto se puede despejar n de la ecuación .

1.7 MUESTREO ALEATORIO SIMPLE CON REEMPLAZO.

En cosas de producción, tiene sentido hacer un MAS con reemplazo, en este caso se está hablando de una distribución de frecuencias, y para sacar la y de la muestra, se tiene que tomar en cuenta las veces que aparece en la muestra la i-ésima unidad, la cual puede ser $1,2,3,...,n$ veces, entonces se tiene que:

$$y = t_1 y_1 + t_2 y_2 + \ldots\ldots + t_N y_N = \sum_{i=1}^{N} t_i y_i$$

Suponiendo que t_i es el número de veces que y_i aparece en la muestra.

De esto se obtiene que la media muestral es:

$$\bar{y} = \frac{y}{n}$$
$$= \frac{1}{n}\sum_{i=1}^{N} t_i y_i$$

Y lo mismo para la varianza, se tendría que calcular de la forma:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} t_i (y_i - \bar{y})^2\right]$$

En este caso, también se cumple que la media muestral es un estimador insesgado de la media poblacional, esto es: $E(\bar{y}) = \bar{Y}$, debido a que la esperanza y la varianza de t_i son:

$E(t_i) = \frac{n}{N}$ (Esperanza).

$Var(t_i) = \frac{n}{N}\left(1 - \frac{1}{N}\right)$ (Varianza).

Tomando en cuenta que las variables t_i, tienen una distribución binomial con parámetros n y $\frac{1}{N}$, para cada unidad muestreada, se tiene que la covarianza entre t_i y t_j es:

$$Cov(t_i, t_j) = -n\left(\frac{1}{N}\right)\left(\frac{1}{N}\right)$$
$$= -\frac{n}{N^2}$$

Debido a que las variables t_i tienen una distribución multinomial. Obteniéndose que para el muestreo con reemplazo, la varianza de la media muestral es:

$$Var(\bar{y}_{CR}) = \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \frac{n(N-1)}{N^2} - 2\sum_{i<j}^{N} y_i y_j \frac{n}{N^2}\right]$$
$$= \frac{1}{nN}\sum_{i=1}^{N}(y_i - \bar{Y})^2$$
$$= \frac{\sigma^2}{n}$$
$$= \frac{N-1}{N}\frac{S^2}{n}$$

El subsímbolo CR es para hacer claro el tipo de muestreo (con reemplazo).

Como ya se vio, la varianza de la media muestral sin reemplazo es:

$$Var(\bar{y}_{SR}) = \frac{N-n}{nN}S^2$$

Nótese que en lo único que varían es que en el numerador en vez de uno (en el caso de muestreo con reemplazo), se tiene una n (en el caso sin reemplazo), esto indica que únicamente varían en los elementos totales, y se puede decir que:

$$Var(\bar{y}_{CR}) = kVar(\bar{y}_{SR}) \qquad con\ k = \frac{N-1}{N-n}$$

Ahora bien, como no se conoce S^2, para obtener la varianza de la media muestral con reemplazo, entonces se estima y se tiene:

$$\hat{Var}(\bar{y}_{CR}) = v(\bar{y}_{CR}) = \frac{N-1}{nN}s^2 \quad \text{donde} \quad s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Una estimación de la población total ($\hat{Y}$) es: $\hat{Y} = N\bar{y}_{CR}$

La varianza de esta población total es:

$$Var(\hat{Y}) = N^2 \frac{(N-1)}{nN} S^2$$
$$= \frac{N(N-1)}{n} S^2$$

pero como no se conoce S^2 entonces se estima y se obtiene:

$$\hat{Var}(\hat{Y}_{CR}) = v(\hat{Y}_{CR}) = \frac{N-1}{nN}s^2$$

1.8 TAMAÑO DE MUESTRA.

La decisión de tomar un tamaño de muestra específico es muy importante, ya que si se toma un tamaño de muestra muy grande, se van a gastar muchos recursos como son tiempo y dinero, y por el otro lado, si se toma un tamaño de muestra pequeño, éste tal vez no de la precisión deseada, y no van a ser conscientes a la hora que se quiera hacer inferencias para la población.

Existen algunas cosas que se necesitan conocer para calcular el tamaño de la muestra y éstas son:

1. Conocer características de la población, si no se conocen se pueden estimar.
2. Saber lo que el investigador quiere, ya sea la precisión del resultado (explicar que se espera), o la confianza que se desea tener (por lo general es 90%, 95% o 99%).
3. Tener las fórmulas que relacionen a n, con algún dato conocido.
4. Si se tiene que en la población existen subdivisiones y éstas tienen límites de error distintos, entonces el tamaño de muestra total será la suma del tamaño de muestra para cada subdivisión.
5. Tener en cuenta que el tamaño de muestra implica costo, tanto en personas que van a llevar a cabo las encuestas como en el material que se va a ocupar.
6. Tomar en cuenta un porcentaje de no respuesta.

1.8.1 TAMAÑO DE MUESTRA PARA PROPORCIONES.

Se tiene que para el caso de proporciones las unidades o pertenecen a la clase C o a su complemento, cuando se vio esto se tenía que la precisión deseada era:

$$d = t\sqrt{\frac{(N-n)PQ}{(N-1)n}}$$

donde:

d es la precisión deseada y por lo tanto se propone.

t es la confianza que se quiere y también se propone, es el cuantil $1-\frac{\alpha}{2}$ de la distribución normal.

p es la proporción empírica de lo que se busca.

Cuando el tamaño de la muestra n es chico, se toma como si t tuviera una distribución t de student con n grados de libertad; en otro caso, se toma como si t tuviera una distribución normal con parámetros (0,1).

De ésta formula se puede sacar el tamaño de muestra n ya que no se conoce P y Q, pero se estiman con p y q, p es la proporción empírica de lo que se busca, y si no se conoce entonces se propone a $p = 0.5$ y $q = 0.5$ ya que éste es el caso menos favorable.

Como los intervalos anteriores no son exactos, debido a que no se conocen P y Q, entonces se toman los siguientes intervalos como una aproximación, que es una corrección por continuidad:

$$p \pm t^{1-\alpha/2}\sqrt{\frac{pq(N-n)}{N(n-1)} + \frac{1}{2n}}$$

Para encontrar los límites para el número total en la clase C en la población, únicamente se multiplica por N, y con esto ya se obtienen los límites de confianza para la población total.

Si de la fórmula de precisión se despeja n, se tiene que:

$$n = \frac{\frac{t^2 PQ}{d^2}}{1 + \frac{1}{N}\left(\frac{t^2 PQ}{d^2} - 1\right)}$$

Para simplificar un poco esta fórmula sea $n_0' = \frac{t^2 PQ}{d^2}$ y se tiene:

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

Si no se conoce a P y Q se les puede estimar con p y q y si no se tiene ni idea de las estimaciones, entonces el tamaño muestral en una distribución binomial con $p = \frac{1}{2}$ y $q = \frac{1}{2}$ se obtiene:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{, donde} \quad n_0 = \frac{t^2 pq}{d^2} = \frac{pq}{Var}$$

Y $Var = \frac{pq}{n_0}$ es la varianza deseada de la proporción de muestra.

Pero si se conoce algo de p entonces la muestra va a empezar a disminuir.

Se sabe que la precisión no es tan sensible a los cambios como la confianza y si se quiere ser más preciso entonces el tamaño de la muestra aumenta, en caso contrario, entonces el tamaño de la muestra es pequeño.

Si se quiere controlar el error relativo r se sustituye rp por d, así se tiene que $d = rp$ donde p es la proporción y d es la precisión y se obtiene que el tamaño de la muestra es:

$$n_0 = \frac{t^2}{r^2}\frac{q}{p} \quad \text{y} \quad n = \frac{n_0}{1+\frac{n_0}{N}}$$

1.8.2 TAMAÑO DE MUESTRA PARA MEDIAS.

Para obtener el tamaño de muestra para medias, se tiene que:

$$\underbrace{\bar{y} \pm t\sqrt{\frac{1-f}{n}S^2}}_{d=\text{precisión}}$$

De la precisión se tiene que:

$d = t\sqrt{\frac{N-n}{Nn}S^2}$, despejando de esta fórmula a n :

$$n_0 = \frac{\frac{t^2}{d^2}S^2}{1+\frac{t^2S^2}{d^2}} \quad \text{y} \quad n = \frac{n_0}{1+\frac{n_0}{N}}$$

Si se supone que el tamaño poblacional tiende a infinito, entonces se tiene que:

$$n_0 = \frac{t^2}{d^2}S^2$$

y como $Var = \frac{d^2}{t^2} \Rightarrow n_0 = \frac{S^2}{Var}$,

pero en este caso como no se puede maximizar, entonces lo que se hace es dar estimadores, y se tiene que:

$$\hat{n}_0 = \frac{t^2}{d^2}s^2$$

$$\hat{n}_0 = \frac{s^2}{\hat{Var}}$$

Y en ambos casos se mantiene el hecho de que $n = \dfrac{n_0}{1+\dfrac{n_0}{N}}$

Si se quiere controlar el error relativo (r), entonces se tiene que $d = r\overline{Y}$

y como $n_0 = \dfrac{t^2}{d^2}S^2$ se sustituye a d, y se tiene que:

$$n_0 = \frac{t^2}{r^2}\left(\frac{S}{\overline{Y}}\right)^2$$

Y como el coeficiente de variación es $C.V. = \dfrac{S}{\overline{Y}}$, entonces se estima a n_0, y se tiene:

$$\hat{n}_0 = \frac{t^2}{r^2}\left(\hat{C}.V.\right)^2 \quad \text{y} \quad n = \frac{n_0}{1+\dfrac{n_0}{N}}$$

El coeficiente de variación anticipado es el que se desea obtener y es:

$$(CV_A)^2 = \frac{r^2}{t^2}$$

El coeficiente de variación poblacional es cuántas veces cabe la media en la desviación estándar y es:

$$(CV_P)^2 = \frac{S^2}{\overline{Y}^2}$$

Y si se pide el tamaño de muestra cuando se da el coeficiente de variación anticipado y el coeficiente de variación poblacional, entonces se tiene que:

$$n_0 = \frac{(CV_P)^2}{(CV_A)^2}$$

Ahora que ya se tiene el tamaño muestral lo que falta es considerar la no respuesta, así el tamaño de muestra verdadero va a ser:

$n_v = n + m(n) = n(1+m)$ donde m es el porcentaje de no respuesta.

Aquí, el porcentaje de no respuesta se dio en forma lineal.

También, se puede considerar la no respuesta en forma geométrica, en este caso, el tamaño de muestra verdadero queda de la forma:

$n_v = \frac{n}{1-m}$, donde m es el porcentaje de no respuesta y en este caso, éste va a ir creciendo rápidamente.

La no respuesta sólo se considera cuando se tiene que entrevistar a personas, o cuando se hace un estudio en donde hay posibilidades de no poder hacer el estudio completo, ya que por otra parte si el investigador saca la información de un archivo es casi imposible que no encuentre toda la información de dicho archivo.

Cuando se está considerando la no respuesta no es lo mismo que sustituir las personas que no contestaron, sino que es para que la confianza, el error relativo y el coeficiente de variación no empeoren, en el caso en el que todas las personas contestaran, entonces lo que pasaría es que se está mejorando las estimaciones.

1.9 EJERCICIOS PROPUESTOS.

1.- De una población con 33 millones de habitantes se ha obtenido una muestra de 10,000. En ella 4,000 se han clasificado como población activa, y de estos, 40 se encuentran en situación de desempleo. Se pide:

a) Estimar el porcentaje de población activa. Estimar también el número de personas activas que se encuentran en situación de desempleo. Calcular el error estándar y el coeficiente de variación en ambas estimaciones. Hallar los intervalos de confianza con riesgo del 3 por mil.

b) ¿Cuántas personas de todas las edades sería necesario incluir en una muestra para estimar la tasa de actividad en España con un error estándar E = 0.02 y una probabilidad del 95%? Del último censo se sabe que en el país hay un 39% de activos.

2.- Mediante muestreo aleatorio simple se trata de estimar la proporción y el total de piezas correctas producidas en un proceso industrial en el que se fabrican un total de 6,000 unidades. Una muestra piloto ha suministrado 1/3 de piezas defectuosas. Se pide:

a) Hallar el tamaño de muestra necesario para que el error estándar sea de una décima al estimar la proporción de piezas correctas producidas en el proceso industrial. Hallar también el tamaño de muestra necesario para que el error relativo de muestreo sea de 20% en la misma estimación.

3.- Mediante muestreo aleatorio simple se obtiene una muestra de tamaño 50 procedente de una población de 750 unidades. Al medir una característica X sobre los elementos de la muestra se obtienen los siguientes datos:

$$\sum_{i=1}^{50} X_i = 454 \qquad \text{y} \quad \sum_{i=1}^{50} X_i^2 = 4306$$

a) Estimar la media y el total de la característica X para la población, así como su error estándar y su coeficiente de variación.

b) Responder a las preguntas del apartado anterior para muestreo aleatorio simple con reemplazo comentando y comparando resultados.

4.- De una población con $N = 100$ unidades se ha extraído una muestra aleatoria simple de tamaño $n = 8$, siendo los datos de una variable X medida sobre ella los siguientes: (25, 32, 28, 35, 26, 34, 30, 28).

a) Basándose en esta muestra estimar la media y el total de la población, así como su error estándar y su coeficiente de variación.

5.- Una muestra aleatoria simple de tamaño $n = 600$ procedente de una población de $N = 15,000$ unidades presenta los siguientes datos para una variable X:

$$\sum_{i=1}^{600} X_i = 2946 \quad \text{y} \quad \hat{S}^2 = 7.06$$

a) Hallar intervalos de confianza al 95% para el total y la media poblacional de X admitiendo normalidad para la distribución de los estimadores.
b) Tomando la muestra anterior como muestra piloto ¿qué tamaño de muestra será necesario para cometer un error estándar de 1,000 unidades al estimar el total de la población anterior?
c) ¿Qué tamaño de muestra será necesario para cometer un coeficiente de variación del 15%?

6.- En una muestra aleatoria simple de 200 obtenida de una población de 2,000 colegios, 120 de estos estuvieron a favor de una propuesta, 57 en contra y 23 se abstuvieron de opinar.

a) Estimar los límites de confianza al 95% para el número de colegios en la población que favorecieron la propuesta.

7.- En una provincia de 300,000 votantes se desea averiguar con confianza del 95% el número de votos que obtendrá un determinado partido político. Para ello, se selecciona una muestra de 5,000 votantes, mediante muestreo aleatorio simple sin reemplazo y se les hace una encuesta. 625 de los encuestados contestaron que estaban dispuestos a votar a dicho partido.

a) Determinar el intervalo de confianza para el número total de votos que obtendrá el partido en las próximas elecciones.

8.- Del total de 5,000 colegios de una comunidad se quiere determinar, con nivel de confianza del 90%, el porcentaje de los que utilizan íntegramente textos de una determinada editorial. Para ello se eligió una muestra aleatoria simple de 220. ¿Cuál es el intervalo de confianza?

9.- Se ha detectado que una marca de aceite para automóviles está causando problemas en los motores, por lo que hay que retirarla de la venta. En una gran ciudad donde existen 3,000 gasolineras, se desea estimar el número total de ellas que vende dicha marca. Para ello se extrajo una muestra aleatoria extraída con reemplazamiento de 150 gasolineras 65 de las cuales entre sus productos ofrecían la marca de aceite que presentaba problemas.

a) Determinar el intervalo de confianza con un nivel de confianza del 95%.

10.- Un examen de estadística realizado a 1,000 alumnos de una facultad consistía en 10 problemas que se puntuaban con 1 si era totalmente correcto o con 0 si tenía algún fallo. Para estimar la nota media de la totalidad de los alumnos, se corrigió una muestra aleatoria de 100 exámenes elegidos sin reemplazamiento con los siguientes resultados:

Número de problemas correctos	0	1	2	3	4	5	6	7	8	9	10
Número de exámenes	2	6	9	8	11	18	15	13	8	6	4

a) Calcular la media poblacional así como la varianza muestral.
b) Estimar el coeficiente de variación.
c) Hallar el intervalo de confianza con nivel de confianza del 95%.

CAPÍTULO 2:

MUESTREO ALEATORIO ESTRATIFICADO.

2.1 MUESTREO ALEATORIO ESTRATIFICADO.

El muestreo estratificado es otro método para muestrear, éste se aplica cuando existe una heterogeneidad en los datos y así se puede reducir dicha heterogeneidad y aumentar la precisión, por lo tanto, si una población tiene características diferentes, y se divide la población de acuerdo a dichas características aplicando a ésta el muestreo aleatorio estratificado, dará un mejor estimador a que si se usa el muestreo aleatorio simple.

El muestreo aleatorio estratificado (M.A.E.), es cuando se parte la población original en subdivisiones y cada una de éstas subdivisiones va ha ser estudiada por separado. Hay que tener en cuenta que estas subdivisiones son excluyentes, ya que la suma de éstas, dan la población original. A cada subdivisión se le denomina estrato. En general, los estratos son heterogéneos entre sí, pero dentro de cada estrato, tienen características similares.

Para poder aplicar el M.A.E. se debe conocer el tamaño de cada estrato y cada estrato se trabaja de manera independiente con el M.A.S., para obtener las estimaciones, aunque el método de estimación las unirá en forma global.

Principalmente el M.A.E. se utiliza para disminuir las varianzas de las estimaciones de la muestra, aunque también es útil debido a que se pueden formar los estratos para utilizar diferentes métodos y procedimientos dentro de cada uno de ellos.

Otra ventaja de utilizar el M.A.E. es porque los estratos pueden establecerse debido a que las subpoblaciones dentro de ellos también se definen como dominios de estudio.

Dominio es una parte de la población para la que se han planeado estimaciones separadas en el diseño de la muestra. Se llama dominio a cualquier subdivisión acerca de la cual se planea la encuesta para proporcionar información numérica de precisión conocida.

NOTACIÓN:

N_h = Tamaño poblacional del estrato ***h***.

$N = \sum_{h=1}^{L} N_h$ = Número total de elementos poblacionales.

n_h = Tamaño muestral del estrato ***h***.

$n = \sum_{h=1}^{L} n_h$ = Número de elementos en la muestra completa.

y_{hi} = valor obtenido para la i-ésima unidad que pertenece al estrato ***h***.

$W_h = \frac{N_h}{N}$ = Ponderación o peso del estrato ***h***. Nótese que: $\sum_{h=1}^{L} W_h = 1$

$f_h = \frac{n_h}{N_h}$ = Fracción de muestreo en el estrato ***h***.

$\overline{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$ = Media verdadera del estrato ***h***.

$\overline{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$ = Media muestral del estrato ***h***.

$$S_h^2 = \frac{\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2}{N_h - 1} = \text{Varianza verdadera en el estrato } h.$$

$$s_h^2 = \frac{\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2}{n_h - 1} = \text{Varianza muestral en el estrato } h.$$

La media estratificada es:

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N} = \sum_{h=1}^{L} W_h \bar{y}_h \quad \text{donde } h = 1,.....,L \ .$$

La media muestral es:

$$\bar{y} = \frac{\sum_{h=1}^{L} n_h \bar{y}_h}{n}$$

Nótese que cuando $\frac{n_h}{n} = \frac{N_h}{N}$ o $\frac{n_h}{N_h} = \frac{n}{N}$ o $f_h = f$ para cada estrato, la media muestral coincide con la media estratificada. Cuando esto sucede se le conoce como **reparto** o **asignación proporcional**.

Algunos teoremas importantes son:

- Si en cada estrato la estimación muestral $\bar{y}_h$ es una estimación insesgada, entonces $\bar{y}_{st}$ es una estimación insesgada de la media de la población $\bar{Y}$ (i.e. $E(\bar{y}_{st}) = \bar{Y}$).[1]

- Si las muestras se extraen independientemente en los diferentes estratos,

  $Var(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 Var(\bar{y}_h)$

  donde $Var(\bar{y}_h)$ es la varianza de $\bar{y}_h$ sobre muestras repetidas del estrato h.[2]

---

[1] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[2] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

- Para el muestreo aleatorio estratificado, la varianza de la estimación $\bar{y}_{st}$ es:

$$Var(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} \quad [3]$$
$$= \sum_{h=1}^{L} \frac{W_h^2 S_h^2 (1 - f_h)}{n_h}$$

- Si las fracciones de muestreo ($f_h = n_h / N_h$) son despreciables en todos los estratos (la población total es infinita),

$$Var(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} \frac{N_h^2 S_h^2}{n_h} \quad [4]$$
$$= \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}$$

- Si se tiene asignación proporcional, i.e. $\frac{N_h}{N} = \frac{n_h}{n}$, entonces la varianza es:

$$Var(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{N_h}{N}\frac{S_h^2}{n}\left(\frac{N-n}{N}\right) \quad [5]$$
$$= \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2$$

- Si el muestreo es proporcional y las varianzas en todos los estratos tienen el mismo valor, S_w^2, entonces:

$$Var(\bar{y}_{st}) = \left(\frac{N-n}{nN}\right) S_w^2 \quad [6]$$

- Si $\hat{Y}_{st} = N\bar{y}_{st}$ es la estimación del total de la población Y, entonces:

$$Var(\hat{Y}_{st}) = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} \quad [7]$$

---

[3] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[4] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[5] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[6] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[7] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

Y una estimación de ésta varianza es:

$$\hat{Var}(\hat{\bar{Y}}_{st}) = v(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} s_h^2 .$$

- Con el M.A.E. una estimación insesgada de la varianza de $\bar{y}_{st}$ es:

$$\hat{Var}(\bar{y}_{st}) = v(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{n_h} s_h^2 .$$

Donde $s_h^2 = \sum_{i=1}^{n_h} \frac{(y_{hi} - \bar{y}_h)^2}{n_h - 1}$ [8]

[8] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

2.2 LÍMITES DE INTERVALOS DE CONFIANZA

Los límites de intervalos de confianza para la media muestral estratificada, se obtienen de $\Pr\left(-t < \frac{\bar{y}_{st} - \mu}{\sigma_{\bar{y}_{st}}} < t\right) = (1-\alpha) * 100\ \%$ de donde se despeja $\bar{y}_{st}$, y se obtiene: $\bar{y}_{st} \pm t\sqrt{Var(\bar{y}_{st})}$.

Una estimación que comúnmente se utiliza es:

$$\bar{y}_{st} \pm t\sqrt{s^2(\bar{y}_{st})} = \bar{y}_{st} \pm \frac{t}{N}\sqrt{\sum_{h=1}^{L} \frac{N_h(N_h - n_h)s_h^2}{n_h}}$$

Por otra parte los límites para el intervalo de confianza para el estimador del total poblacional son: $\hat{Y} \pm t\sqrt{Var(\hat{Y})}$ o el estimador $\hat{Y} \pm t\sqrt{s^2(\hat{Y})}$ el cuál también se puede obtener como: $N\bar{y}_{st} \pm Nt\sqrt{s^2(\bar{y}_{st})} = N\left(\bar{y}_{st} \pm t\sqrt{s^2(\bar{y}_{st})}\right)$.

Estas fórmulas suponen que $\bar{y}_{st}$ está normalmente distribuida y que la varianza está bien determinada, así t es el cuantil $1-\alpha$ de la distribución normal.

2.3 MUESTREO DESPROPORCIONADO O AFIJACIÓN OPTIMA.

Si se permite que haya un número fijo de elementos $n = \sum_{h=1}^{L} n_h$, la varianza de $\bar{y}_{st}$ puede hacerse mínima si la razón de muestreo dentro de cada estrato se hace proporcional a la desviación estándar dentro del estrato, lo cuál se obtiene si los L valores de f_h se escogen de manera que:

$f_h = \frac{n_h}{N_h} = kS_h$, donde k es una constante de proporcionalidad.

Si en lugar de fijar n, se especifica la varianza, entonces la afijación de las n_h a los diversos estratos, de acuerdo con la fórmula anterior, proporcionará una varianza especificada para la n más pequeña. Por lo general, la varianza disminuye conforme el tamaño de muestra va aumentando.

Ahora se supone que el costo por elemento en los diferentes estratos es de C_h el cuál es proporcional al tamaño de la muestra. Lo que se quiere es minimizar el costo para un valor específico de la varianza $Var(\bar{y}_{st})$, o minimizar dicha varianza para un costo específico. Para cualquiera de estos dos casos, la asignación óptima se alcanza cuando las razones de muestreo dentro de los estratos se hacen directamente proporcionales a las desviaciones estándar dentro de los estratos o inversamente proporcionales a las raíces cuadradas de los costos por elemento dentro de los estratos. Dicho de otra manera :

$f_h = \frac{n_h}{N_h} = K\frac{S_h}{\sqrt{C_h}}$, donde K es una constante de proporcionalidad.

Para poder obtener una asignación óptima, se debe de conocer la función de costo, y la más simple es de forma lineal como:

$C = C_0 + \sum_{h=1}^{L} C_h n_h$, donde C es el costo y C_0 es un costo fijo.

A continuación se mencionarán algunos teoremas importantes:

- En el M.A.E., con una función de costo $C = C_0 + \sum_{h=1}^{L} C_h n_h$, la varianza de la media estimada $\bar{y}_{st}$ es un mínimo para un costo específico C y el costo es un mínimo para una varianza específica $Var(\bar{y}_{st})$, donde n_h es proporcional a $\frac{W_h S_h}{\sqrt{C_h}}$.[9]

Del teorema anterior se puede decir que:

1.- La varianza mínima para un costo fijo es:

$$Var_{min}(\bar{y}_{st}) = \sum_{h=1}^{L} \sqrt{C_h} W_h S_h - \sum_{h=1}^{L} \frac{N_h S_h^2}{N^2}.$$

Y el tamaño de muestra cuando la varianza mínima y el costo fijo es:

$$n = \frac{(C - C_0)\sum_{h=1}^{L} \frac{N_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \sqrt{C_h} N_h S_h}$$

2.- El costo mínimo para una varianza fija es:

$$C_{min} = C_0 + \sum_{h=1}^{L} W_h S_h \sqrt{C_h}$$

Y el tamaño de muestra cuando la varianza es fija y el costo mínimo es:

$$n = \frac{\left(\sum_{h=1}^{L} \sqrt{C_h} W_h S_h\right)\left(\sum_{h=1}^{L} \frac{W_h S_h}{\sqrt{C_h}}\right)}{Var(\bar{y}_{st}) + \sum_{h=1}^{L} \frac{W_h S_h^2}{N}}$$

[9] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

Cuando no existe una especificación de reparto entonces el tamaño de muestra se puede obtener mediante la fórmula:

$$n = \frac{n_0}{1 + \frac{1}{NV}\sum_{h=1}^{L} W_h s_h^2},$$

donde $n_0 = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{w_h V}$, y

$V = Var(\bar{y}_{st})$ deseada (es fija), o si no, se puede tomar como:

$$V = \left(\frac{d}{t}\right)^2 = \left(\frac{precisión}{confianza}\right)^2.$$

Cuando el reparto es proporcional (i.e. $n_h = W_h n$), entonces se tiene que el tamaño de muestra se puede obtener mediante la fórmula:

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

donde $n_0 = \frac{\sum_{h=1}^{L} W_h s_h^2}{V}$ y

$V = Var(\bar{y}_{st})$ deseada (es fija), o si no, se puede tomar como:

$$V = \left(\frac{d}{t}\right)^2 = \left(\frac{precisión}{confianza}\right)^2.$$

Cuando el reparto es el óptimo, i.e. $n_h = \frac{n W_h s_h}{\sum_{h=1}^{L} W_h s_h}$, entonces el tamaño de muestra se puede calcular mediante la fórmula:

$$n = \frac{\left(\sum_{h=1}^{L} W_h s_h\right)^2}{V + \frac{\sum_{h=1}^{L} W_h s_h^2}{N}}, \text{ donde}$$

$V = Var(\bar{y}_{st})$ deseada (es fija), o si no, se puede tomar como:

$$V = \left(\frac{d}{t}\right)^2 = \left(\frac{precisión}{confianza}\right)^2.$$

- En el M.A.E. la $Var(\bar{y}_{st})$ se minimiza para un tamaño de muestra total fijo n si:

$$n_h = \frac{nW_hS_h}{\sum_{h=1}^{L} W_hS_h} = \frac{nN_hS_h}{\sum_{h=1}^{L} N_hS_h},$$

a ésta asignación se le conoce como la asignación de Neyman.[10]

Y una fórmula para la varianza mínima, que es la varianza óptima, con n fija es:

$$Var_{min}(\bar{y}_{st}) = \frac{\left(\sum_{h=1}^{L} W_hS_h\right)^2}{n} - \frac{\sum_{h=1}^{L} W_hS_h^2}{N}.$$

Al principio del capítulo se mencionó que por lo general el M.A.E. da como resultado una varianza menor que si se aplicara el M.A.S., pero en el caso en que los valores de n_h están lejos del óptimo, el M.A.E. puede tener una varianza más grande, y por esta razón a continuación se mencionará un teorema que hace una comparación con ambos tipos de muestreo.

- Si se ignoran los términos en $\frac{1}{N_h}$, relativamente a la unidad,

$Var_{opt} \leq Var_{prop} \leq Var_{M.A.S.}$ (Esto sólo sucede cuando N es muy grande).

donde la asignación óptima, que es la mínima, es para n fijo, o sea, con $n_h \propto N_hS_h$

[10] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

Donde:

$$Var_{opt}(\bar{y}_{st}) = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{n} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N}.$$

$$Var_{prop}(\bar{y}_{st}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 \quad \text{con } n_h = nW_h .$$

$$Var_{M.A.S}(\bar{y}) = \frac{1-f}{n} S^2.$$[11]

[11] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

2.4 MUESTREO ALEATORIO ESTRATIFICADO PARA PROPORCIONES.

En este caso, si la población se divide en diferentes estratos, entonces la muestra va a guardar la proporción de cada uno de los estratos, esto va a permitir tener un muestreo representativo. En las muestras proporcionales, la fracción de muestreo en cada estrato se hace igual a la fracción de muestreo para la población completa (i.e. $\frac{n_h}{N_h} = \frac{n}{N}$ para todos los estratos). Además, se representa a todos los estratos en la muestra con las mismas razones (i.e. $\frac{n_h}{n} = \frac{N_h}{N} W_h$ para todos los estratos).

De lo anterior se puede decir que el muestreo proporcional se da cuando $\frac{n_h}{n} = \frac{N_h}{N}$.

Con esto se puede estimar la media de la población muestral la cual es:

$$\bar{y}_{prop} = \frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

Como los factores $1-f_h$ son constantes entonces la varianza de una muestra proporcional se puede obtener con la fórmula:

$$Var(\bar{y}_{prop}) = \frac{1-f}{n}\left[\sum_{h=1}^{L} W_h {s_h}^2\right]$$
$$= \frac{1-f}{n} {s_w}^2$$

El valor muestral ${s_w}^2$ estima a la varianza por elemento dentro del estrato ${S_w}^2$.

La varianza del total para una muestra proporcional es:

$$Var(y) = (1-f)\sum_{h=1}^{L} \frac{n_h}{n_h - 1}\left[\sum_{h=1}^{L} y^2{}_{hi} - \frac{{y_h}^2}{n_h}\right]$$

En general, se obtienen solamente ganancias pequeñas a partir de un muestreo proporcional de elemento, porque las variables disponibles para estratificación, tales como edad y sexo, no separan a la población en estratos muy homogéneos. Las razones por las cuales se utiliza el muestreo proporcional es porque a menudo produce ganancias modestas en varianzas reducidas, además, las varianzas que se obtengan de una muestra no estratificada serán siempre mayores y por lo general son fáciles y simples, pero es común que las pequeñas ganancias que se obtienen con el muestreo proporcional, se habrían ganado con un pequeño aumento en el tamaño de la muestra utilizando el M.A.S..

En este tipo de muestreo se quiere estimar la proporción de unidades en la población que pertenecen a una clase definida *C*.

NOTACIÓN:

$P_h = \frac{A_h}{N_h}$ = Proporción poblacional de individuos que pertenecen al estrato *h* y tienen la característica *C*.

$p_h = \frac{a_h}{n_h}$ = proporción muestral de individuos que pertenecen al estrato *h* y tienen la característica *C*.

$p_{st} = \sum_{h=1}^{L} W_h p_h$ = La estimación de la proporción en la población total.

(Esta fórmula vendría a ser $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$ en el M.A.E.).

$S_h^2 = \frac{N_h P_h Q_h}{N_h - 1}$ = Varianza por estrato.

Como por lo general no se conoce la proporción de la población lo que se hace es estimarla con la proporción muestral.

Un teorema importante relacionado con este tema es:

- Con el M.A.E., la varianza de p_{st} es:

$$Var(p_{st}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2(N_h - n_h)}{N_h - 1}\frac{P_h Q_h}{n_h}.$$ [12]

- Cuando se pueda ignorar el factor de corrección por finitud, entonces:

$$Var(p_{st}) = \sum_{h=1}^{L}\frac{W_h^2 P_h Q_h}{n_h}.$$ [13]

- Con asignación proporcional la varianza de p_{st} es:

$$Var(p_{st}) = \frac{N-n}{N}\left(\frac{1}{nN}\right)\sum_{h=1}^{L}\frac{N_h^2 P_h Q_h}{N_h - 1}.$$ [14]

- Con asignación óptima (mínima)

$n_h = \dfrac{nW_h S_h}{\sum_{h=1}^{L} W_h S_h} = \dfrac{nN_h S_h}{\sum_{h=1}^{L} N_h S_h}$, la varianza mínima de p_{st} es:

$$Var_{\min}(\bar{p}_{st}) = \frac{\left(\sum_{h=1}^{L} W_h\sqrt{P_h Q_h}\right)^2}{n}.$$ [15]

---

[12] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[13] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[14] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[15] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

Para conocer el tamaño de muestra por estrato el reparto óptimo se va a dar con:

$$n_h = n \frac{W_h \sqrt{\frac{N_h P_h Q_h}{(N_h - 1)\, C_h}}}{\sum_{h=1}^{L} W_h \sqrt{\frac{N_h P_h Q_h}{(N_h - 1)\, C_h}}}$$

donde n se calcula con las fórmulas mencionadas anteriormente dependiendo si se quiere un costo mínimo o una varianza mínima.

El tamaño de muestra para proporciones con asignación óptima se obtiene mediante la fórmula:

$$n \doteq \frac{\left(\sum_{h=1}^{L} W_h \sqrt{p_h q_h}\right)^2}{V + \frac{1}{N} \sum_{h=1}^{L} W_h p_h q_h},$$

donde la varianza se puede dar en términos de precisión y confianza.

El tamaño de muestra para proporciones con asignación proporcional es:

$$n \doteq \frac{\sum_{h=1}^{L} W_h p_h q_h}{V + \frac{1}{N} \sum_{h=1}^{L} W_h p_h q_h},$$

aquí también la varianza se puede dar en términos de precisión y confianza.

2.5 EJERCICIOS PROPUESTOS.

1.- Se desea conocer el número de Km que recorren al mes los conductores de una ciudad. Comentar las ventajas e inconvenientes que puedan tener las siguientes variables tomadas para la estratificación: Sexo, edad, distancia al trabajo, tipo de trabajo.
Proponer otros criterios para la estratificación.

2.- Una población de tamaño 1,000 está dividida en tres estratos para los que se conocen los siguientes datos: $\sigma_1 = 4, \sigma_2 = 12, \sigma_3 = 80$, $W_1 = 0.6$, $W_2 = 0.3$, $W_3 = 0.1$.
Se pide:

a) Determinar el tamaño de muestra que con afijación proporcional da una varianza del estimador de la media igual a 5, considerando muestreo sin reemplazo. Realizar las respectivas afijaciones proporcionales.
b) ¿Qué resultados se obtendrían con afijación de mínima varianza? Realizar las respectivas afijaciones de mínima varianza.
c) Determinar el tamaño de muestra para afijación óptima con costos $C_1 = 1{,}000$, $C_2 = 1{,}200$, $C_3 = 2{,}000$, considerando muestreo sin reemplazo.

3.- Determinar el tamaño de n de la muestra estratificada que con afijación de mínima varianza produzca la misma precisión que una muestra aleatoria simple (no estratificada) de tamaño n`, para estimar la proporción $P$ de una cierta clase en la población. Suponer en ambos casos muestreo con reemplazo y aplicar el resultado a los datos de la tabla con n`=1,000.

	Estratos		
	I	II	III
W_h	0.2	0.3	0.5
P_h	0.5	0.6	0.4

4.- Supongamos conocidos los siguientes datos de una población dividida en tres estratos $S_1^2 = 9$, $S_2^2 = 225$, $S_3^2 = 1600$, $N_1 = 1000$, $N_2 = 600$, $N_3 = 200$, $C_1 = 1000$, $C_2 = 1200$, $C_3 = 2000$. Se pide lo siguiente:
Determinar el costo de una muestra estratificada que proporciona un coeficiente de variación de 5% para estimar la media considerando afijación de mínima varianza. Se sabe que $\overline{X} = 22$ y que la función de costo es lineal.

5.- Se van a muestrear las familias de un pueblo para estimar la cantidad promedio de bienes por familia que se pueden convertir en dinero efectivo rápidamente. Las familias se estratifican en un estrato de renta alta y otro de renta baja. Se piensa que una casa en el estrato de renta alta tiene cerca de 9 veces más bienes que una casa en el estrato de renta baja, y se espera que S_h sea proporcional a la raíz cuadrada de la media del estrato. Se sabe que existen 4,000

familias en el estrato de renta alta y 20,000 familias en el estrato de renta baja. Se pide:

a) ¿Cómo se distribuiría de forma óptima entre los dos estratos una muestra de 1,000 familias extraida de la población?

6.- Una empresa de publicidad quiere estimar la proporción de hogares en un municipio donde se ve cierto programa televisivo. El municipio tiene en total 310 hogares y es dividido en tres estratos. Una muestra estratificada de n=40 hogares se selecciona con afijación proporcional. Estimar la proporción de hogares en el municipio donde se ve el programa televisivo estimando el error estándar y el coeficiente de variación cometido. Datos:

Estratos	Tamaños muestrales	No. De hogares donde se ve el programa	$\hat{P}_k$
1	n_1=20	16	0.80
2	n_2=8	2	0.25
3	n_3=12	6	0.50

7.-En una población constituida por 21,500 individuos se agrupan estos en cuatro estratos, en los que:

Número de individuos del estrato	4,000	2,500	10,000	5,000
Varianza del estrato	2.5^2	8^2	4^2	6^2

a) Determinar el tamaño muestral en cada estrato, en base al criterio de afijación de mínima varianza, si el tamaño muestral total es n=100.

b) Si , elegidas las muestras según el criterio de afijación anterior, se obtienen las medias muestrales en cada estrato resultando ser:
$\bar{x}_1 = 2,\ \bar{x}_2 = 3,\ \bar{x}_3 = 1.5\ \bar{x}_4 = 2.4$
Determinar el valor estimado de la media poblacional.

8.- Una población de 10,000 individuos está dividida en dos estratos. Los valores de W_i y S_i son como sigue:

Estrato	W_i	S_i
1	0.4	10
2	0.6	20

a) ¿Cuál ha de ser el tamaño de la muestra si se quiere que el error muestral para la media sea igual a 1.5, utilizando afijación proporcional?

9.- Una población de 15,000 habitantes se divide en tres estratos caracterizados por la edad de los individuos con el fin de estimar el número medio de horas que dedican semanalmente a realizar algún tipo de ejercicio físico. Se estudia una muestra de 300 individuos que proporciona los siguientes valores sobre la media y varianza en los diferentes estratos.

Estratos	N_i	n_i	$\bar{x}_i$	S_i^2
1	6,000	133	2.5	1.3
2	5,500	96	3.5	0.8
3	3,500	71	5.5	1.1

a) Calcular el estimador de la media.
b) Calcular la estimación de la varianza.
c) Calcular los tamaños de muestra en cada estrato para estimar la media según el criterio de afijación proporcional.
d) Calcular el intervalo de confianza con nivel de confianza del 90% para estimar el número medio de horas que dedican semanalmente a realizar algún tipo de ejercicio físico.

10.- Una población está dividida en varios estratos, con distinta varianza e importancia económica, cuyos valores se reflejan en la tabla siguiente:

N_i	S_i^2
500	1,750
1,000	850
1,500	1,250
2,000	1,000
5,000	500

a) Si se obtuviera una muestra de tamaño 500, ¿De qué tamaños deberían ser las submuestras por estratos, según los criterios de afijación proporcional y afijación de mínima varianza?.
b) De que tamaño debería ser la muestra total según el criterio de afijación proporcional, si se desea estimar la media de la población con un error máximo de 5 unidades y nivel de significación del 5%.

CAPÍTULO 3:

MUESTREO

SISTEMÁTICO.

3.1 MUESTREO SISTEMÁTICO.

Este método de muestreo también se le conoce como muestreo seudoaleatorio y es quizá el más conocido y usado, ya que es más fácil sacar una muestra sin cometer errores además de ahorrar tiempo. Consiste en tomar una muestra aleatoria de N unidades seleccionando n lo que se hace es tomar una unidad (r) aleatoriamente entre los primeros k elementos (nótese que $k = \frac{N}{n}$) y después se toma una unidad sumándole al primer elemento seleccionado (r), k unidades. Así se tendrá que las unidades de la muestra serán: r, $r+k$, $r+2k$, $r+3k$,.....*etc.*.

Si k no es un número entero, lo que se hace es escoger k de tal manera que N sea mayor que nk, pero menor que $\frac{(n+1)}{k}$, o bien, se reducen los listados a exactamente nk elementos antes de iniciar. Pero por lo general lo que se hace es considerar a la lista como un círculo, es decir, después del último elemento, sigue el primero, y así hasta alcanzar las n unidades deseadas.

Algunas veces en vez de tomar el primer elemento aleatoriamente se toma como $\frac{(k+1)}{2}$ si k es impar o $\frac{k}{2}$ si k es par. El muestreo sistemático se usa generalmente en poblaciones en las cuales la numeración de las unidades es efectivamente aleatoria.

Muchas veces se utiliza conjuntamente con la estratificación o con conglomerados[1], ya que las particiones de k, 2k, 3k, ... generan conglomerados de igual tamaño y así se selecciona un elemento de cada conglomerado, así la muestra

[1] Un conglomerado es un grupo que tiene características semejantes y contiene mas de un elemento de la población. Ver pagina 71, capítulo 4.

se reparte más uniformemente sobre la población, donde cada elemento tiene la probabilidad de $\frac{1}{k}$ de ser elegido. Una muestra sistemática es una m.a.s. de una unidad conglomerada, tomada en una población de *k* unidades conglomeradas. Tanto el muestreo aleatorio estratificado como el sistemático son mucho más efectivos que el muestreo aleatorio simple, pero el muestreo sistemático tiene menos precisión que el aleatorio estratificado, aunque esto no siempre se cumple.

En un estrato de *ck* unidades el error relativo es menor que $\frac{1}{c}$.

La media de la muestra es una estimación insesgada de la media de la población, si el tamaño de la muestra se ha fijado en *n* $\left(\text{i.e. } E(\bar{y}_{sy}) = \bar{Y}\right)$.

Sea $\bar{y}_{sy}$ = Media de una muestra sistemática.

y_{ij} = *j-ésimo* miembro de la *i-ésima* muestra sistemática *j=1, 2, ...n; i=1, 2,...,k.*

$\bar{y}_{i\bullet}$ = media de la *i-ésima* muestra.

$\bar{y}_{\bullet j}$ = La media del estrato.

Los teoremas que dicen cual es la varianza de la media de una muestra sistemática, que se aplican a cualquier clase de muestreo conglomerado son:

- La varianza de la media de una muestra sistemática es:

  $$Var(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S^2_{wsy}$$

  donde $S^2_{wsy} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\bullet})^2$ = varianza entre las unidades que se encuentran dentro de la misma muestra sistemática.[2]

[2] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

- La media de una muestra sistemática es más precisa que la media de una muestra aleatoria simple sí y solamente sí:

  $S^2_{wsy} > S^2$.[3]

  En otras palabras, esto dice que el muestreo sistemático es más preciso que el M.A.S. si la varianza dentro de las muestras sistemáticas es mayor que la varianza de la población total.

- Otra forma para la varianza de la media de una muestra sistemática es:

$$Var(\bar{y}_{sy}) = \frac{S^2}{n}\left(\frac{N-1}{N}\right)\left[1+(n-1)\rho_w\right]$$

  donde ρ_w es el coeficiente de correlación entre pares de unidades que están en la misma muestra sistemática y se define como:

$$\rho_w = \frac{E\left[(y_{ij}-\bar{Y})(y_{iu}-\bar{Y})\right]}{E(y_{ij}-\bar{Y})^2}$$

  o bien: $\rho_w = \frac{2}{(n-1)(N-1)S^2}\sum_{i=1}^{k}\sum_{j<u}(y_{ij}-\bar{Y})(y_{iu}-\bar{Y})$.[4]

- La varianza de la media de una muestra sistemática que se da en términos de una muestra aleatoria estratificada es:

$$Var(\bar{y}_{sy}) = \frac{S^2_{wst}}{n}\left(\frac{N-n}{N}\right)\left[1+(n-1)\rho_{wst}\right]$$

  donde $S^2_{wst} = \frac{1}{n(k-1)}\sum_{j=1}^{n}\sum_{i=1}^{k}(y_{ij}-\bar{y}_u)^2$ = varianza entre unidades comprendidas en el mismo estrato.[5]

---

[3] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[4] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[5] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

$\rho_{wst} = \frac{E(y_{ij} - \bar{y}_{\bullet j})(y_{iu} - \bar{y}_{\bullet u})}{E(y_{ij} - \bar{y}_{\bullet j})^2}$ = Correlación entre las desviaciones de los pares de características de las medias de los estratos que están en la misma muestra sistemática.

$$\rho_{wst} = \frac{2}{n(n-1)(k-1)} \sum_{i=1}^{k} \sum_{j<u} \frac{(y_{ij} - \bar{y}_j)(y_{iu} - \bar{y}_u)}{S_{wst}^2}$$

- Una muestra sistemática tiene la misma precisión que la correspondiente muestra aleatoria estratificada con una unidad por estrato sí $\rho_{\omega\sigma\tau} = 0$, debido a que:

$$Var(\bar{y}_{st}) = \left(\frac{N-n}{N}\right)\frac{S_{wst}^2}{n}.$$ [6]

El teorema a continuación muestra como en promedio, la varianza del muestreo sistemático es equivalente al muestreo aleatorio simple.

- Considérense todas las *N!* poblaciones finitas las cuales están formadas por las N! permutaciones de cualquier conjunto de números $y_1, y_2, y_3, \ldots, y_N$. Entonces el promedio sobre estas poblaciones finitas, $E(Var(\bar{y}_{sy})) = Var(\bar{y})$. Nótese que la varianza del M.A.S. es la misma para todas las permutaciones. [7]

  Sea ε el promedio sobre todas las poblaciones finitas que se pueden sacar de la superpoblación (población infinita que tiene ciertas propiedades).

---

[6] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.
[7] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

- Si las varianzas y_i $(i=1, 2, ...,N)$ se sacan al azar de una superpoblación en la cual $\varepsilon y_i=\mu$, $\varepsilon(y_i-\mu)(y_j-\mu)=0$ $(i\neq j)$ $\varepsilon(y_i-\mu)^2=\sigma_i^2$ (i.e. que todos los y_i tienen la misma media μ y que la varianza σ_i^2 puede cambiar de punto a punto en la serie.) entonces: $\varepsilon\, Var(\bar{y}_{sy}) = \varepsilon\, Var(\bar{y})$.[8]

Hay varias alternativas:

1.- Aleatorizar el orden de la población, pero esto no es práctico.

2.- Introducir la selección aleatoria como simple, o dentro de estratos.

3.- Cambiar varias veces el arranque aleatorio.

4.- La selección replicada de c muestras diferentes.

Si la población presenta una tendencia lineal entonces se tiene que la varianza en una muestra aleatoria simple es:

$$Var(\bar{y}) = \frac{N-n}{N}\frac{S^2}{n}$$
$$= \frac{n(k-1)}{N}\frac{N(N+1)}{12n} \quad \text{debido a que } S^2 = \frac{N(N+1)}{12}.$$
$$= \frac{(k-1)(N+1)}{12}$$

Para obtener la varianza estratificada lo que se hace es que la varianza dentro de los estratos es $S_w^2 = \frac{k(k+1)}{12}$, por lo tanto se tiene:

$$Var(\bar{y}_{st}) = \frac{N-n}{N}\frac{S_w^2}{n}$$
$$= \frac{n(k-1)}{nk}\frac{k(k+1)}{12n}.$$
$$= \frac{(k^2-1)}{12n}$$

[8] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980:

Y la varianza de una muestra sistemática es:

$$Var(\bar{y}_{sy}) = \frac{1}{k}\sum_{u=1}^{k}(\bar{y}_u - \bar{Y})^2 = \frac{k^2-1}{12}$$, debido a que $\sum_{u=1}^{k}(\bar{y}_u - \bar{Y})^2 = \frac{k(k^2-1)}{12}$.

Con esto se deduce que: $Var(\bar{y}_{st}) \leq Var(\bar{y}_{sy}) \leq Var(\bar{y})$

Esto indica que el muestreo sistemático resulta mucho más efectivo que el M.A.S. pero menos efectivo que el M.A.E.

En este caso, en que se presente una tendencia lineal, se puede mejorar, utilizando una muestra centralmente localizada o bien cambiando la estimación de una media no ponderada a una media ponderada.

Hay tres métodos mejores que el muestreo sistemático ordinario en presencia de una tendencia lineal o parabólica y son:

1. Muestreo sistemático de correcciones extremas (examinado por Yates). Éste fue extendido por Bellhouse y Rao.
2. Muestreo sistemático equilibrado (que fue propuesto por Sethi, pero nombrado así por Murthy).
3. Muestreo sistemático modificado (propuesto por Singh).

Si la población presenta una tendencia periódica, entonces la efectividad del muestreo sistemático ordinario depende del valor de k, ya que éste puede ser igual, o un múltiplo entero del periodo, y es aquí cuando se presenta el caso menos favorable. Y el caso más favorable es cuando k es un múltiplo impar del semiperiodo. (Un ejemplo son las ventas que se tienen cada semana en una tienda,

por lo tanto en este caso, no es aconsejable obtener la muestra de lo que se vendió en un día de la semana específico, sino que es mejor irle cambiando).

Hay ocasiones en las que se espera que dos observaciones consecutivas sean positivamente correlacionadas, ya que en poblaciones naturales, se pueden inducir cambios lentos al recorrer la serie. La correlación depende de la distancias entre las dos observaciones, si se supone que las dos observaciones son y_i y y_j, entonces su separación es *i-j* , y su correlación disminuye al crecer la distancia. Para saber si se puede aplicar este modelo a una población, lo que se hace es calcular el conjunto de correlaciones ρ_u para pares de atributos que distan u unidades, y graficar esta correlación respecto a u, llamada correlograma, el cual es muy irregular si se tiene una población finita, y esto dificulta la deducción de resultados. Entonces lo que se hace es el promedio de una serie de poblaciones finitas tomadas al azar de una suprapoblación infinita a la que se le aplica el modelo. La varianza promedio para el muestreo sistemático se denota con:

$$\varepsilon Var(\bar{y}_{sy}) = \varepsilon\ E(\bar{y}_{sy} - \bar{Y})^2$$

Aquí no se puede decir cual de los dos muestreos es mejor, si el estratificado o el sistemático, ya que esto depende de los valores de k.

Este teorema que a continuación se menciona supone que el correlograma es cóncavo hacia arriba.

- Si $\varepsilon(y_i)=\mu$, $\varepsilon(y_i-\mu)^2=\sigma^2$, $\varepsilon(y_i-\mu)(y_{i+u}-\mu)=\rho_u\sigma^2$ con $\rho_u \geq \rho_v \geq 0$,
  Siempre que $u < v$ y además
  $\partial_u^2 = \rho_{u+1} + \rho_{u-1} - 2\rho_u \geq 0$ $[u = 2,3,\ldots,(kn-2)]$, entonces
  $\varepsilon\ Var(\bar{y}_{sy}) \leq \varepsilon\, Var(\bar{y}_{st}) \leq \varepsilon\, Var(\bar{y})$.
  para cualquier tamaño de muestra.
  Además, a menos que $\partial_u^2 = 0$, $u=2, 3, \ldots, (kn-2)$,

$\varepsilon\ Var(\bar{y}_{sy}) < \varepsilon\ Var(\bar{y}_{st})$.[9]

Debido a la concavidad, la muestra estratificada pierde más en precisión, cuando la distancia es menor que *k*, de lo que gana, cuando la distancia excede a *k*.

Parece que en general, una muestra aleatoria estratificada con estratos de tamaño *2k* y dos unidades por estrato, es mejor en precisión que una muestra aleatoria simple. Y una muestra sistemática, es un poco mejor en precisión que una muestra con estratos de tamaño *k* y una unidad por estrato.

Es imposible construir una varianza estimada que no tenga sesgo si hay variación periódica. Pero se puede construir un modelo matemático que represente adecuadamente el tipo de variación que se tiene, el cual va a depender del juicio del muestreador. Estos modelos, se aplican a poblaciones en las cuales y_i se compone de una tendencia más una componente aleatoria, la cual tiene las características que $\varepsilon(e_i) = 0$; $\varepsilon(e_i^2) = \sigma^2$; $\varepsilon(e_i e_j) = 0$, (dicho de otra manera $y_i = \mu_i + e_i$) . Así una fórmula propuesta para la varianza estimada se denomina insesgada si es sin sesgo para todas las poblaciones finitas que se puedan extraer de la superpoblación esto es: $\varepsilon\, E(s_{sy}^2) = \varepsilon\, V_{sy}$.

Algunos modelos simples son:

a) ***Población en orden aleatorio***.- se aplica cuando se esta seguro de que el orden es aleatorio. La formula de varianza es la misma que en m.a.s. y no se presenta sesgo alguno.

μ_i = constante.

$$s_{syl}^2 = \frac{N-n}{Nn}\frac{\sum_{i=1}^{N}(y_i - \bar{y}_{sy})^2}{n-1}.$$

[9] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

b) ***Solamente efectos de estratificación***.- aquí la media es constante dentro de cada estrato de k unidades. la estimación s_{sy2}^2 que se basa en la diferencia sucesiva del cuadrado medio, no es insesgada.

μ_i = constante.

$$s_{sy2}^2 = \frac{N-n}{Nn}\frac{\sum_{i=rk+1}^{rk+k}(y_i - y_{i+k})^2}{2(n-1)}.$$

c) ***Tendencia lineal***.- esta estimación se basa en términos cuadráticos sucesivos en la secuencia y_i .

$\mu_i = \mu + \beta_i$.

$$s_{sy3}^2 = \frac{N-n}{N}\frac{n'}{n^2}\frac{\sum_{i=1}^{n-2}(y_i - 2y_{i+k} + y_{i+2k})^2}{6(n-2)}.$$

El termino $\frac{n'}{n^2}$ se puede reemplazar por $\frac{1}{n}$ en el caso en que n sea muy pequeña.

La formula cuadrática (s_{sy2}^2) es un poco mejor que la basada en diferencias sucesivas (s_{sy3}^2), pero ambas producen sobreestimaciones.

Si lo que se quiere son estimaciones separadas para cada estrato o si se quiere usar fracciones de muestreo desiguales, lo que se hace es sacar una muestra sistemática separada dentro de cada estrato con puntos iniciales determinados independientemente. Si $\bar{y}_{syh}$ es la media de la muestra sistemática en el estrato h, entonces:

$\hat{\bar{y}}_{stsy} = \sum_{h=1}^{L} W_h \bar{y}_{syh}$ = la estimación $\bar{Y}$ de la media de población .

$Var(\bar{y}_{stsy}) = \sum_{h=1}^{L} W_h^2 Var(\bar{y}_{syh})$ = la varianza de la estimación de la media de la población.

Cuando el número de estratos es grande, se utiliza una estimación basada en el método de los estratos contraídos, la cual da como resultado:

$\hat{Var}(\bar{y}_{stsy}) = \sum W_h^2 (\bar{y}_{syh} - \bar{y}_{syj})^2$, esta suma se extiende sobre los pares de estratos.

Se tienen dos tipos de muestra sistemática en dos dimensiones: el de la rejilla cuadrada o muestra alineada y la muestra no alineada, la cual es mejor que la rejilla cuadrada y el m.a.e. y consta de seleccionar primero la unidad superior izquierda aleatoriamente, y otros dos números aleatorios determinan las coordenadas horizontales, después se requieren otros dos para fijar las coordenadas verticales de las unidades restantes en la primera fila de estratos, y el intervalo constante *k* fija las posiciones de todos los puntos.

3.2 EJERCICIOS PROPUESTOS.

1.- Dada la población siguiente:

u_i	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
X_i	1	3	5	2	4	6	2	7	3

se desea obtener una muestra sistemática de tamaño 3 (1 en 3).

a) Determinar el espacio muestral y las probabilidades asociadas a las muestras posibles para este tipo de muestreo.
b) Calcular las varianzas de los estimadores insesgados del total y de la media.
c) Seleccionar la muestra más precisa.

2.- Dada la población siguiente:

u_i	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
X_i	1	3	5	2	4	6	2	7

se realiza muestreo sistemático 1 en 2. Se pide:

a) Calcular las varianzas de los estimadores insesgados del total y de la media. Utilizar adicionalmente la relación entre muestreo sistemático y estratificado.
b) Seleccionar la muestra más precisa.

3.- Considérese una población cuyos elementos se están ordenados del siguiente modo X = (0, 1, 2, 3, 4, 5, 6, 7, 8), y tómese una reordenación alternativa Y = (1, 4, 5, 8, 6, 0, 3, 2, 7)
Mediante muestreo sistemático se toman muestras de tamaño 3.

a) Determinar en cada caso las posibles muestras que se podrían obtener y la media de las mismas.
b) Calcular en ambos casos la varianza de la media muestral.

4.- Un vecindario tiene tres comunidades compactas, que consisten respectivamente, de personas de descendencia anglosajona, polaca e italiana. Hay un directorio actualizado. En él, las personas de una casa están enlistadas en el siguiente orden: marido, esposa, hijos (por edades), otros. Las casas son enlistadas en orden a lo largo de las calles. El número promedio de personas por casa es de cinco.
La elección es entre una muestra sistemática de cada quinta persona en el directorio y una muestra aleatoria simple del 20%. ¿Para cuáles de las siguientes variables se espera que la muestra sistemática sea más precisa? a) Proporción de personas de descendencia polaca, b) proporción de varones, c) Proporción de hijos.

5.- En un directorio de 13 casas de una calle las personas están distribuidas hogar a hogar como sigue:

1	2	3	4	5	6	7	8	9	10	11	12	13
M	M	M	M	M	M	M	M	M	M	M	M	M
F	F	F	F	F	F	F	F	F	F	F	F	F
F	f	m		m	f	f	m	m	m	f	f	
M	m	f		m	m	f	f		f	m		
F	f			f		m						

M=varón adulto, F=mujer adulta, m=hijo varón, f=hija.

Se realiza muestreo sistemático de una de cada 5 personas (muestreo 1 en 5), numerando los elementos de la población por columnas hacia abajo y luego yendo a la parte superior de la siguiente columna (se empieza por la primera columna de la izquierda). Se pide lo siguiente:

a) Calcular el valor del coeficiente de correlación ρ_{wst} y hallar la varianza del estimador de la proporción de varones adultos en la población .

6.- Las 36 viviendas de una calle numeradas del 1 al 36 se ordenan alfabéticamente en un archivo de acuerdo con el apellido del jefe de familia. Las viviendas cuyo jefe de familia es extranjero son las que tienen los números 2, 5-7, 11-13, 15-16, 20-22, 25-26, 28 y 30-34.

a) Calcular las varianzas del estimador insesgado de la media.
b) Comparar la precisión de una muestra sistemática 1 en 4 con una muestra aleatoria simple del mismo tamaño para estimar la proporción de viviendas en las cuales el jefe de familia es extranjero.

7.- Una población de 360 viviendas (numeradas de 1 a 360) en Baltimore se ordena alfabéticamente en un archivo de acuerdo con el apellido del jefe de familia. Viviendas en las que el jefe no es blanco ocurren en los números siguientes: 28, 31-33, 36-41, 44, 45, 47, 55, 56, 58, 68, 69, 82, 83, 85, 86, 89-94, 98, 99, 101, 107-110, 114, 154, 156, 178, 223, 224, 296,298-300, 302-304, 306-323, 325-331, 333, 335-339, 341, 342. (Las viviendas de no blancos muestran algún "agrupamiento" debido a la asociación entre apellido y color).

a) Compare la precisión de una muestra sistemática 1 en 8 con una muestra aleatoria simple del mismo tamaño para estimar la proporción de viviendas en las cuales el jefe de familia no es blanco.

CAPÍTULO 4:

MUESTREO POR

CONGLOMERADOS

4.1 MUESTREO POR CONGLOMERADOS.

En ocasiones, el investigador se encuentra con el problema de no poder contar o disponer de un marco poblacional, ya que las unidades últimas de muestreo no se conocen completamente, o bien en el caso de que se conozcan, resulta muy costoso por dinero o tiempo a invertir el elaborar el marco correspondiente. En estos casos lo conveniente es dividir a la población en grupos o segmentos de población denominados conglomerados, los cuales se esperaría que replicaran aproximadamente el mismo comportamiento poblacional en cuanto a las características que se están estudiando y que naturalmente deberán constar de más de un elemento de la población.

Así, la principal característica en el muestreo aleatorio por conglomerados es que las unidades de muestreo en la primera o primeras etapas son grupos de elementos y contienen más de uno. Cada una de las unidades últimas de muestreo deberán de estar identificadas unívocamente, es decir, deberán pertenecer a uno y solamente un conglomerado; por ejemplo, en una ciudad en la que se desee conocer el número de viviendas rentadas y el número de viviendas propias, sería demasiado costoso formar el marco poblacional de todas las viviendas que se localizan en la ciudad, entonces, lo que el investigador podría sugerir es dividir la ciudad en conglomerados, los cuales podrían ser las colonias, y con éstas se podría obtener el número de las viviendas que son rentadas o que son propias, en este caso éstas serían las unidades últimas de muestreo.

En el muestreo de conglomerados se tiene una ventaja y una desventaja, en la primera, el costo por elemento es menor (esto se debe a que la localización es más fácil) y en la segunda, la varianza por elemento es mayor, debido al procedimiento de selección de los elementos de la población.

En la implantación del muestreo aleatorio de conglomerados se tiene que éstos pueden ser de igual tamaño o bien desiguales, los primeros se dan principalmente en los procesos de fabricación de jeringas, tornillos, comida enlatada, entre otras cosas y los segundos se observan en los casos de la madre naturaleza como por ejemplo en papas, zanahorias, naranjas, por sólo citar algunos ejemplos, cabe decir que éstos últimos son los más comunes y frecuentes.

4.2 SELECCIÓN ALEATORIA DE CONGLOMERADOS.

Si se supone que la muestra contiene a conglòmerados que se seleccionan de los A conglomerados que tiene la población total y además si se supone que cada conglomerado de los a que existen en la muestra tiene B elementos (conglomerados de igual tamaño), los cuales se incluyen en la muestra (por lo que $n = a * B$), entonces se tiene que la probabilidad de selección de los N elementos de la población (donde $N = A * B$) es :

$$\frac{a}{A}=\frac{a}{A}\frac{B}{B}=\frac{n}{N}=f.$$

Sea a = el número de conglomerados de la muestra.

A = el número de conglomerados en la población total.

B = el número de elementos en el conglomerado.

$y_{\alpha\beta}$ = el valor observado para el *β-ésimo* elemento del *α-ésimo* conglomerado.

$$\sum_{\beta=1}^{B} y_{\alpha\beta} = y_{\alpha}$$

$\bar{y}_{\alpha} = \frac{1}{B} y_{\alpha}$ = media muestral del *α-ésimo* conglomerado.

$$\bar{y} = \frac{1}{a}\sum_{\alpha=1}^{a} \bar{y}_{\alpha}$$

$$\sum_{\alpha=1}^{a} \bar{y}_{\alpha} = \frac{1}{B}\sum_{\alpha=1}^{a} y_{\alpha} = \frac{y}{B}$$

$y = \sum_{\alpha=1}^{a} y_{\alpha}$ = total muestral.

$\bar{y} = \frac{1}{n}\sum_{\alpha=1}^{a} y_{\alpha} = \frac{y}{n}$ = media muestral.

Se puede estimar la media de la población ($\bar{Y}$) a partir de la media muestral de los n elementos, la cual también es la media de las a medias por conglomerado siendo:

$$\bar{y} = \frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{aB}\sum_{\alpha=1}^{a}\sum_{\beta=1}^{B} y_{\alpha\beta} = \frac{1}{aB}\sum_{\alpha=1}^{a} y_\alpha = \frac{1}{a}\sum_{\alpha=1}^{a} \bar{y}_\alpha$$

$\bar{Y}_\alpha$ = media poblacional del *α-ésimo* conglomerado.

$$\bar{Y}_\alpha = \frac{1}{B}(Y_{\alpha 1} + Y_{\alpha 2} + \ldots + Y_{\alpha B}) = \frac{1}{B}\sum_{\beta=1}^{B} Y_{\alpha\beta} = \frac{Y_\alpha}{B}$$

$\bar{Y}$ = media del total de la población.

$$\bar{Y} = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{1}{AB}\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B} Y_{\alpha\beta} = \frac{1}{A}\left[\frac{1}{B}\sum_{\alpha=1}^{A} Y_\alpha\right] = \frac{1}{A}\sum_{\alpha=1}^{A} \bar{Y}_\alpha .$$

Es de señalar que el subíndice α en la población total va de 1 a A y en la muestra se cuenta de 1 a a, y del mismo modo el subíndice β va de 1 a B para los conglomerados de igual tamaño, ya que B indica el número de elementos que contiene cada conglomerado.

La media muestral es un estimador insesgado de la media poblacional ($\bar{Y}$), dicho de otra manera: $E(\bar{y}) = \bar{Y}$

La varianza de la media muestral se puede definir como:

$$Var(\bar{y}) = 1 - f\frac{s_a^2}{a}$$

donde $s_a^2 = \frac{1}{a-1}\sum_{\alpha=1}^{a}(\bar{y}_\alpha - \bar{y})^2$ = varianza entre medias por unidad.

Esta varianza es directamente proporcional a la varianza entre unidades de muestreo e inversamente proporcional al número de unidades de muestreo. Nótese que primero se obtiene la varianza de los conglomerados y con esto, se obtiene la varianza de la media de la muestra.

Un equivalente a esta varianza sería:

$$Var(\bar{y}) = \frac{1-f}{a}\frac{s_y^2}{B^2} \text{ donde } s_y^2 = B^2 s_a^2 = \left[\frac{1}{(a-1)}\left(\sum_{\alpha=1}^{a} y_\alpha^2 - \frac{y^2}{a}\right)\right]$$

por demostrar que: $s_y^2 = B^2 s_a^2 = \left[\frac{1}{(a-1)}\left(\sum_{\alpha=1}^{a} y_\alpha^2 - \frac{y^2}{a}\right)\right]$

donde $s_a^2 = \frac{1}{a-1}\sum_{\alpha=1}^{a}(\bar{y}_\alpha - \bar{y})^2$

p.d. $B^2\left(\frac{1}{a-1}\right)\sum_{\alpha=1}^{a}(\bar{y}_\alpha - \bar{y})^2 = \left[\frac{1}{(a-1)}\left(\sum_{\alpha=1}^{a} y_\alpha^2 - \frac{y^2}{a}\right)\right]$

multiplicando ambos lados por $(a - 1)$, se tiene:

$$B^2\sum_{\alpha=1}^{a}(\bar{y}_\alpha - \bar{y})^2 = \sum_{\alpha=1}^{a} y_\alpha^2 - \frac{y^2}{a}$$

Desarrollando el primer lado de la igualdad, se tiene:

$$B^2\sum_{\alpha=1}^{a}(\bar{y}_\alpha - \bar{y})^2 = B^2\sum_{\alpha=1}^{a}(\bar{y}_\alpha^2 - 2\bar{y}_\alpha\bar{y} + \bar{y}^2)$$

$$= B^2\sum_{\alpha=1}^{a}\bar{y}_\alpha^2 - 2B^2\bar{y}\sum_{\alpha=1}^{a}\bar{y}_\alpha + B^2\sum_{\alpha=1}^{a}\bar{y}^2$$

$$= B^2\sum_{\alpha=1}^{a}\bar{y}_\alpha^2 - 2B^2\bar{y}\left(\frac{y}{B}\right) + B^2\bar{y}^2 a\,, \qquad \text{ya que } \sum_{\alpha=1}^{a}\bar{y}_\alpha = \frac{y}{B}.$$

$$= \sum_{\alpha=1}^{a} (B\bar{y}_\alpha)^2 - 2B\bar{y}y + B^2 a\bar{y}^2$$

$$= \sum_{\alpha=1}^{a} \left(B\left(\frac{y_\alpha}{B}\right)\right)^2 - B\bar{y}(2y - Ba\bar{y}) , \qquad \text{ya que } \bar{y}_\alpha = \frac{1}{B} y_\alpha$$

$$= \sum_{\alpha=1}^{a} y_\alpha^2 - B\bar{y}\left(2y - (n)\left(\frac{y}{n}\right)\right), \qquad \text{ya que } n = aB \quad \text{y que } \bar{y} = \frac{y}{n}$$

$$= \sum_{\alpha=1}^{a} y_\alpha^2 - B\bar{y}(2y - y) = \sum_{\alpha=1}^{a} y_\alpha^2 - B\bar{y}y$$

$$= \sum_{\alpha=1}^{a} y_\alpha^2 - \left(\frac{n}{a}\right)\left(\frac{y}{n}\right)y, \qquad \text{ya que } B = \frac{n}{a} \quad \text{y que } \bar{y} = \frac{y}{n}$$

$$= \sum_{\alpha=1}^{a} y_\alpha^2 - \frac{y^2}{a}$$

por lo tanto:

$$B^2 \sum_{\alpha=1}^{a} (\bar{y}_\alpha - \bar{y})^2 = \sum_{\alpha=1}^{a} y_\alpha^2 - \frac{y^2}{a}$$

El estimador simple e insesgado del total de la población Y es:

$$\hat{Y} = N\bar{y}$$
$$= N\frac{y}{n}$$
$$= \frac{N}{n} y \qquad \text{ya que } f = \frac{n}{N}$$
$$= \frac{y}{f}$$

por lo tanto $\hat{Y} = \frac{y}{f}$

Y su error estándar es:

$$ee(\hat{Y}) = \frac{N\sqrt{(1-f)\,s_a^{\,2}}}{\sqrt{a}}$$ ya que:

$$\begin{aligned} ee(\hat{Y}) &= \sqrt{\text{var}(\hat{Y})} \\ &= \sqrt{\text{var}(N\bar{y})} \\ &= \sqrt{N^2\,\text{var}(\bar{y})} \\ &= N\sqrt{\text{var}(\bar{y})} \\ &= N\sqrt{1-f\frac{s_a^2}{a}} \\ &= \frac{N\sqrt{(1-f)s_a^2}}{\sqrt{a}} \end{aligned}$$

4.3 MUESTREO DE CONGLOMERADOS EN UNA ETAPA: CONGLOMERADOS DE TAMAÑOS DESIGUALES.

Como ya se mencionó, la mayoría de las veces las unidades conglomerados son de diferentes tamaños, así pues, se verán a continuación 2 métodos para estimar el total de la población *Y*.

A) MUESTREO ALEATORIO SIMPLE DE CONGLOMERADOS: ESTIMACION INSESGADA.

Sea:

M_i = el número de elementos de la *i-ésima* unidad (conglomerado).

y_{ij} = valor observado para el *j-ésimo* elemento de la *i-ésima* unidad.

$y_i = \sum_{j=1}^{M_i} y_{ij} = M_i \bar{y}_i$ = Total de elementos para la *i-ésima* unidad de conglomerado.

$\bar{y}_i = \frac{y_i}{M_i}$ = Media muestral para la *i-ésima* unidad de conglomerado.

$\bar{Y} = \frac{Y}{N}$ = Media de población por conglomerado.

$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i$ = Estimación insesgada del total de la población *Y*.

Esta estimación no requiere del conocimiento de todos los M_i.

$Var(\hat{Y}) = \frac{N^2(1-f)}{n} \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N-1}$ = Varianza de la estimación del total de la población.

Cuando las medias por elemento varían poco entre unidades, y los M_i varían considerablemente, se tiene que esta estimación es un poco deficiente y la varianza es grande.

B) MUESTREO ALEATORIO SIMPLE DE CONGLOMERADOS: ESTIMACIÓN DE RAZÓN A TAMAÑO.

Sea:

$M_0 = \sum_{i=1}^{N} M_i$ = Número total de elementos de la población.

$\bar{\bar{Y}} = \frac{\sum_{i=1}^{N} y_i}{NM} = \frac{\bar{Y}}{M}$ = Media por elemento.

N = Número de conglomerados.

M = Tamaño del conglomerado.

Si los M_i y por lo tanto, M_0 se conocen, una alternativa es la estimación de razón en donde M_i se toma como la variable auxiliar x_i, entonces se tiene que:

$$\hat{Y}_R = M_0 \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$$

Esta estimación requiere del conocimiento del total M_0, es decir, de todos los M_i.

La varianza de esta media muestral, depende de la variabilidad entre las medias por elemento y si se supone que el número de conglomerados de la muestra es grande, se tiene que es:

$$Var(\hat{Y}_R) \doteq \frac{N^2(1-f)}{n}\frac{\sum_{i=1}^{N}(y_i - M_i\bar{\bar{Y}})^2}{N-1}$$

$$\doteq \frac{N^2(1-f)}{n}\frac{\sum_{i=1}^{N}M_i^2(\bar{y}_i - \bar{\bar{Y}})^2}{N-1}$$

Y por lo general se tiene que : $Var(\hat{Y}_R) < Var(\hat{Y})$.

Si se estima la media de población por elemento, se tiene entonces:

$$\hat{\bar{\bar{Y}}} = \frac{\hat{Y}}{M_0} = \frac{N}{nM_0}\sum_{i-1}^{n} y_i$$

Y si se estima la media muestral por elemento se tiene:

$$\hat{\bar{\bar{Y}}}_R = \frac{\hat{Y}_R}{M_0} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$$

Ésta estimación solo requiere del conocimiento de los M_i que caen dentro de la muestra seleccionada.

4.4 SELECCIÓN CON PROBABILIDADES DESIGUALES Y CON RESTITUCIÓN.

Sea:

$\frac{M_i}{M_0}$= Probabilidad con que se ha elegido la *i-ésima* unidad con restitución.

$M_0 = \sum_{i=1}^{N} M_i$ = Número total de elementos en todas las unidades.

Una estimación insesgada del total de la población Y es:

$\hat{Y} = \frac{M_0}{n}(\bar{y}_1 + \bar{y}_2 + ... + \bar{y}_n)$=Media de las medias por elemento de las unidades.

Por notación: sea $\hat{Y} = \hat{Y}_{pps}$

La varianza de esta estimación insesgada del total de la población es:

$$Var(\hat{Y}_{pps}) = \frac{M_0}{n}\sum_{i=1}^{N} M_i(\bar{y}_i - \bar{\bar{Y}})^2.$$

En ocasiones, el tamaño de M_i no es el número de elementos de la unidad, sino una medida de su extensión que se piensa altamente correlacionada con el total de unidad y_i, por ejemplo el tamaño de un hospital, podría medirse por el número total de camas, o por el número promedio de camas ocupadas en un cierto periodo de tiempo, entonces se considera una medida de tamaño M_i' y una probabilidad correspondiente de selección z_i.

Sea : M_i' = Medida del tamaño de la *i-ésima* unidad.

$z_i = \frac{M_i'}{M_0'}$ = Probabilidad correspondiente de selección.

donde: $M_0' = \sum_{i=1}^{N} M_i'$ y $\sum_{i=1}^{N} z_i = 1$

TEOREMA 1.

- *Si una muestra de n unidades se extrae con las probabilidades z_i y con reemplazo, entonces:*

$$\hat{Y}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i}$$

es una estimación insesgada de Y con varianza:

$$Var(\hat{Y}_{pps}) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2 .^{1}$$

por demostrar que:

$$Var(\hat{Y}_{pps}) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2 \quad \text{si } \hat{Y}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i}$$

Sea t_i = el número de veces que aparece la *i-ésima* unidad en una muestra específica de tamaño n.

Por lo tanto $t_i = 0, 1, 2, \ldots, n$, esto es, tiene una distribución multinomial.

Entonces se tiene que:

$E(t_i) = n z_i$, $\quad Var(t_i) = n z_i (1 - z_i)$ $\quad$ y $\quad Cov(t_i\ t_j) = -n z_i\ z_j$

donde z_i es la probabilidad de que una bola determinada vaya a la caja *i-ésima*, entonces se puede escribir:

$$\hat{Y}_{pps} = \frac{1}{n}\left(t_1\frac{y_1}{z_1} + t_2\frac{y_2}{z_2} + \ldots + t_N\frac{y_N}{z_N}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{N} t_i\frac{y_i}{z_i}$$

Tanto y_i como z_i son números fijos. La única variable aleatoria es t_i, así se tiene que:

---

[1] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

$$E(\hat{Y}_{pps}) = E\left(\frac{1}{n}\sum_{i=1}^{N} t_i \frac{y_i}{z_i}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{N} \frac{y_i}{z_i} E(t_i)$$
$$= \frac{1}{n}\sum_{i=1}^{N} \frac{y_i}{z_i} n z_i$$
$$= \frac{1}{n} n \sum_{i=1}^{N} \frac{y_i}{z_i} z_i$$
$$= \sum_{i=1}^{N} y_i = Y$$

Por lo tanto $E(\hat{Y}_{pps}) = Y$, esto es que $\hat{Y}_{pps}$ es insesgada.

Ahora bien

$$Var(\hat{Y}_{pps}) = Var\left(\frac{1}{n}\sum_{i=1}^{N} t_i \frac{y_i}{z_i}\right)$$
$$= \frac{1}{n^2} Var\left(\sum_{i=1}^{N} t_i \frac{y_i}{z_i}\right)$$
$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} Var\left(t_i \frac{y_i}{z_i}\right) + 2Cov\left(\frac{y_i}{z_i} t_i, \frac{y_j}{z_j} t_j\right)\right]$$
$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} \left(\frac{y_i}{z_i}\right)^2 Var(t_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{y_i}{z_i}\frac{y_j}{z_j} Cov(t_i, t_j)\right]$$

Como $Var(t_i) = n z_i(1 - z_i)$ y $Cov(t_i, t_j) = -n z_i z_j$, entonces se tiene que:

$$Var(\hat{Y}_{pps}) = \frac{1}{n^2}\left[\sum_{i=1}^{N} \left(\frac{y_i}{z_i}\right)^2 n z_i(1 - z_i) - 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{y_i}{z_i}\frac{y_j}{z_j} n z_i z_j\right]$$
$$= \frac{n}{n^2}\left[\sum_{i=1}^{N} \frac{y_i^2(1 - z_i)}{z_i} - 2\sum_{i=1}^{N}\sum_{j>i}^{N} y_i y_j\right]$$
$$= \frac{1}{n}\left[\sum_{i=1}^{N} \frac{y_i^2}{z_i} - \sum_{i=1}^{N} y_i^2 - 2\sum_{i=1}^{N}\sum_{j>i}^{N} y_i y_j\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{N}\frac{y_i^2}{z_i} - 2\sum_{i=1}^{N}\sum_{j>i}^{N} y_i y_j - \sum_{i=1}^{N} y_i^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{N}\frac{y_i^2}{z_i} - Y^2\right] = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2 \text{ ; esto ya que } \sum_{i=1}^{N} z_i = 1$$

Por lo tanto:

$$Var\left(\hat{Y}_{pps}\right) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2$$

TEOREMA 2.

- *Si se extrae una muestra de n unidades, con probabilidad proporcional a z_i con reemplazo, para cualquier $n > 1$, la estimación muestral insesgada de $Var(\hat{Y}_{pps})$ estará dada por:*

$$\hat{Var}(\hat{Y}_{pps}) = \frac{\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - \hat{Y}_{pps}\right)^2}{n(n-1)}$$ [2]

p.d. $\hat{Var}(\hat{Y}_{pps}) = \dfrac{\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - \hat{Y}_{pps}\right)^2}{n(n-1)}$ es una estimación insesgada de $Var(\hat{Y}_{pps})$.

p.d. $E\left(\hat{Var}(\hat{Y}_{pps})\right) = \text{var}(\hat{Y}_{pps})$

$$\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - \hat{Y}_{pps}\right)^2 = \sum_{i=1}^{n}\left(\frac{y_i}{z_i} - Y - (\hat{Y}_{pps} - Y)\right)^2$$

$$= \sum_{i=1}^{n}\left[\left(\frac{y_i}{z_i} - Y\right)^2 - 2\left(\frac{y_i}{z_i} - Y\right)\left(\hat{Y}_{pps} - Y\right) + \left(\hat{Y}_{pps} - Y\right)^2\right]$$

[2] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

$$= \sum_{i=1}^{n}\left(\frac{y_i}{z_i} - Y\right)^2 - 2(\hat{Y}_{pps} - Y)\sum\left(\frac{y_i}{z_i} - Y\right) + n(\hat{Y}_{pps} - Y)^2$$

$$\text{como } \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i} = \hat{Y}_{pps} \Rightarrow \sum_{i=1}^{n}\frac{y_i}{z_i} = n\hat{Y}_{pps}$$

$$= \sum_{i=1}^{n}\left(\frac{y_i}{z_i} - Y\right)^2 - 2(\hat{Y}_{pps} - Y)n(\hat{Y}_{pps} - Y) + n(\hat{Y}_{pps} - Y)^2$$

$$= \sum_{i=1}^{n}\left(\frac{y_i}{z_i} - Y\right)^2 - n(\hat{Y}_{pps} - Y)^2$$

Por lo tanto:

$$\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - \hat{Y}_{pps}\right)^2 = \sum_{i=1}^{n}\left(\frac{y_i}{z_i} - Y\right)^2 - n(\hat{Y}_{pps} - Y)^2$$

Ahora bien, como $\hat{V}ar(\hat{Y}_{pps}) = \dfrac{\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - \hat{Y}_{pps}\right)^2}{n(n-1)}$,

entonces se tiene que:

$$n(n-1)\hat{V}ar(\hat{Y}_{pps}) = \sum_{i=1}^{n}\left(\frac{y_i}{z_i} - \hat{Y}_{pps}\right)^2$$

$$= \sum_{i=1}^{n}\left(\frac{y_i}{z_i} - Y\right)^2 - n(\hat{Y}_{pps} - Y)^2$$

sacando la esperanza a la expresión de arriba, se tiene que:

$$E\left[n(n-1)\hat{V}ar(\hat{Y}_{pps})\right] = E\left[\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - Y\right)^2 - n(\hat{Y}_{pps} - Y)^2\right]$$

Esto implica que:

$$n(n-1)E\left[\hat{V}ar(\hat{Y}_{pps})\right] = E\left[\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - Y\right)^2\right] - nE\left[(\hat{Y}_{pps} - Y)^2\right]$$

Por definición : $Var(\hat{Y}_{pps}) = E\left[\left(\hat{Y}_{pps} - Y\right)^2\right]$.

Por lo tanto se tiene que:

$$n(n-1)E\left[\hat{V}ar(\hat{Y}_{pps})\right] = E\left[\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - Y\right)^2\right] - nVar(\hat{Y}_{pps})$$

Introduciendo la variable aleatoria t_i (la cual se definió antes), se tiene:

$$n(n-1)E\left[\hat{V}ar(\hat{Y}_{pps})\right] = E\left[\sum_{i=1}^{N} t_i\left(\frac{y_i}{z_i} - Y\right)^2\right] - nVar(\hat{Y}_{pps})$$

$$= \sum_{i=1}^{N} E\left[t_i\left(\frac{y_i}{z_i} - Y\right)^2\right] - n\,\mathrm{var}(\hat{Y}_{pps})$$

$$= \sum_{i=1}^{N}\left(\frac{y_i}{z_i} - Y\right)^2 E(t_i) - n\,\mathrm{var}(\hat{Y}_{pps})$$

Y como $E(t_i) = nz_i$, entonces se tiene que:

$$n(n-1)E\left[\hat{V}ar(\hat{Y}_{pps})\right] = \sum_{i=1}^{N}\left(\frac{y_i}{z_i} - Y\right)^2 nz_i - n\,\mathrm{var}(\hat{Y}_{pps})$$

$$= n\sum_{i=1}^{N}\left(\frac{y_i}{z_i} - Y\right)^2 z_i - n\,\mathrm{var}(\hat{Y}_{pps})$$

Y como ya se demostró en el teorema anterior que:

$$Var(\hat{Y}_{pps}) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2 \Rightarrow nVar(\hat{Y}_{pps}) = \sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2$$

Por lo tanto

$$n(n-1)E\left[\hat{V}ar(\hat{Y}_{pps})\right] = nn\,\mathrm{var}(\hat{Y}_{pps}) - n\,\mathrm{var}(\hat{Y}_{pps})$$

$$= n\,\mathrm{var}(\hat{Y}_{pps})[n-1]$$

$$E\left[\hat{V}ar(\hat{Y}_{pps})\right] = \frac{1}{n(n-1)}Var(\hat{Y}_{pps})(n(n-1))$$

$$E\left(\hat{V}ar(\hat{Y}_{pps})\right) = Var(\hat{Y}_{pps})$$

TEOREMA 3.

- *Si se extrae una muestra de n unidades con probabilidades proporcionales al tamaño* $z_i = \frac{M_i}{M_0}$ *y con reemplazo:*

$$\hat{Y}_{pps} = \frac{M_0}{n}\sum_{i=1}^{n}\left(\frac{y_i}{M_i}\right) = \frac{M_0}{n}\sum_{i=1}^{n}(\bar{y}_i) = M_0\bar{\bar{y}}$$

donde $\bar{\bar{y}}$ *es la media no ponderada de las medias de unidades, es una estimación insesgada de Y con varianza:* $Var(\hat{Y}_{pps}) = \frac{M_0}{n}\sum_{i=1}^{N} M_i(\bar{y}_i - \bar{\bar{Y}})^2$

dado que $\bar{y}_i = \frac{y_i}{M_i}$ *y* $\bar{\bar{Y}} = \frac{Y}{M_0}$.[3]

Demostración:

p.d. $Var(\hat{Y}_{pps}) = \frac{M_0}{n}\sum_{i=1}^{N} M_i(\bar{y}_i - \bar{\bar{Y}})^2$ y $E(\bar{\bar{y}}) = Y$

Como $\hat{Y}_{pps} = \frac{M_o}{n}\sum_{i=1}^{n}\left(\frac{y_i}{M_i}\right) = \frac{M_o}{n}\sum_{i=1}^{n}(\bar{y}_i) = M_o\bar{\bar{y}}$

Donde $\bar{\bar{y}}$ = media no ponderada de las medias de unidades.

Se tiene que:

$$\bar{y}_i = \frac{y_i}{M_i} \Rightarrow y_i = \bar{y}_i M_i$$

$$\bar{\bar{Y}} = \frac{Y}{M_o} \Rightarrow Y = M_o\bar{\bar{Y}}$$

$$\hat{Y}_{pps} = M_o\bar{\bar{y}}$$

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{n}\bar{y}_i}{n}$$

Por el teorema 1 se sabe que:

$$Var(\hat{Y}_{pps}) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2$$

[3] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

Como $y_i = \bar{y}_i M_i$ y $Y = M_o \bar{\bar{Y}}$ y $z_i = \frac{M_i}{M_o}$.

$$Var(\hat{Y}_{pps}) = \frac{1}{n}\sum_{i=1}^{N} \frac{M_i}{M_o}\left(\frac{\bar{y}_i M_i}{\frac{M_i}{M_o}} - M_o\bar{\bar{Y}}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{N} \frac{M_i}{M_o}\left(\frac{\bar{y}_i M_i M_o}{M_i} - M_o\bar{\bar{Y}}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{N} \frac{M_i}{M_o}\left(\bar{y}_i M_o - M_o\bar{\bar{Y}}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{N} \frac{M_i}{M_o} M_o^2(\bar{y}_i - \bar{\bar{Y}})^2$$

$$= \frac{1}{n}\sum_{i=1}^{N} M_i M_o(\bar{y}_i - \bar{\bar{Y}})^2$$

$$= \frac{M_o}{n}\sum_{i=1}^{N} M_i(\bar{y}_i - \bar{\bar{Y}})^2$$

Por lo tanto se tiene que:

$$Var(\hat{Y}_{pps}) = \frac{M_o}{n}\sum_{i=1}^{N} M_i(\bar{y}_i - \bar{\bar{Y}})$$

TEOREMA 4.

- *En las condiciones del teorema anterior, una estimación muestral insesgada de* $Var(\hat{Y}_{pps})$ *es:*

$$\hat{Var}(\hat{Y}_{pps}) = M_0^2 \sum_{i=1}^{n} \frac{(\bar{y}_i - \bar{\bar{y}})^2}{n(n-1)}$$

dado que :

$$\hat{Y}_{pps} = M_0\bar{\bar{y}} \quad ; \qquad z_i = \frac{M_i}{M_0} \quad ; \qquad \bar{y}_i = \frac{y_i}{M_i}.$$ [4]

[4] Cochran Gemell William, Técnicas de Muestreo, México CECSA, 1980.

p.d. $\hat{V}ar(\hat{Y}_{pps}) = M_o^2 \sum_{i=1}^{n} \frac{(\bar{y}_i - \bar{\bar{y}})^2}{n(n-1)}$

Del teorema 2 se tiene que:

$$\hat{V}ar(\hat{Y}_{pps}) = \frac{\sum_{i=1}^{n} \left(\frac{y_i}{z_i} - \hat{Y}_{pps} \right)^2}{n(n-1)}$$

Sustituyendo $z_i = \frac{M_i}{M_0}$

Y como se sabe que $\bar{y}_i = \frac{y_i}{M_i} \Rightarrow y_i = \bar{y}_i M_i$

$$\hat{Y}_{pps} = M_o \bar{\bar{y}}$$

$$\hat{V}ar(\hat{Y}_{pps}) = \frac{\sum_{i=1}^{n} \left(\frac{\bar{y}_i M_i}{\frac{M_i}{M_o}} - M_o \bar{\bar{y}} \right)^2}{n(n-1)}$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\bar{y}_i M_i M_o}{M_i} - M_o \bar{\bar{y}} \right)^2$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\bar{y}_i M_o - M_o \bar{\bar{y}} \right)^2$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} M_o^2 \left(\bar{y}_i - \bar{\bar{y}} \right)^2$$

Por lo tanto se tiene que:

$$\hat{V}ar(\hat{Y}_{pps}) = \frac{M_o^2 \sum_{i=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2}{n(n-1)}$$

4.5 EJERCICIOS PROPUESTOS.

1.- Una población está formada por 300 conglomerados de 50 elementos cada uno. Se obtiene una muestra de 5 conglomerados sin reposición y probabilidades iguales. Las proporciones de unidades elementales que pertenecen a una cierta clase en cada uno de los conglomerados son 0.14, 0.20, 0.18, 0.12, 0.16. Se pide:

a) Estimar el total de clase, su error estándar y el coeficiente de variación.

2.- En una muestra aleatoria de 10 viviendas, el número de personas y sus contestaciones afirmativas a una determinada pregunta fueron:

Número de personas	4	2	6	1	5	3	3	8	1	4
Contestaciones afirmativas	2	1	4	1	2	1	2	5	0	3

a) Suponiendo muestreo con reposición estimar la proporción de respuestas afirmativas en la población y su error de muestreo.

3.- Se trata de hacer un control de calidad sobre las piezas de un pedido de 1,000 lotes formados por 40 piezas cada uno. Para ello se extrae una muestra sin reposición de 20 lotes dentro de la cual 9 lotes no tienen piezas defectuosas, 8 lotes tienen una pieza defectuosa y 3 lotes tiene dos piezas defectuosas. Se pide:

a) Estimar el número total de piezas defectuosas en el pedido, su error estándar y el coeficiente de variación.

4.- Se trata de estudiar la superficie de una región montañosa dedicada a la plantación de pinos. La región se divide en 100 zonas disjuntas lo más similares entre sí de tal forma que cada zona contiene plantas de todas las clases que crecen en la región. Se extrae una muestra de 10 zonas con reemplazamiento y con probabilidades proporcionales a sus superficies. Las proporciones de superficie total dedicadas a la plantación de pinos en cada una de las zonas de la muestra son: 0.05, 0.25, 0.10, 0.30, 0.15, 0.25, 0.35, 0.25, 0.10 y 0.20
Proporcione un estimador insesgado de la superficie total de la región dedicada a la plantación de pinos y su error relativo.

5.- Un fabricante de sierras quiere estimar el costo de reparación promedio mensual para las sierras que ha vendido a ciertas industrias. El fabricante no puede obtener un costo de reparación por sierra, pero puede obtener la cantidad total gastada en reparación y el número de sierras que tiene cada industria. El fabricante decide seleccionar una muestra aleatoria simple sin reposición de 20 industrias de entre las 96 a las que ofrece servicio. Los datos de gasto total mensual en reparaciones por industria y el número de sierras por industria se presenta en la tabla siguiente:

Industria	Número de sierras	Costo total reparaciones mensual	Industria	Número de sierras	Costo total reparaciones mensual
1	3	50	11	8	140
2	7	110	12	6	130
3	11	230	13	3	70
4	9	140	14	2	50
5	2	60	15	1	10
6	12	280	16	4	60
7	14	240	17	12	280
8	3	45	18	6	150
9	5	60	19	5	110
10	9	230	20	8	120

a) Estimar el costo promedio de reparación mensual por sierra y su error de muestreo.
b) Estimar la cantidad gastada por las 96 industrias en la reparación de sierras y su error de muestreo.
c) Después de verificar sus registros de ventas el fabricante se percata de que ha vendido un total de 710 sierras a esas industrias. Usando esta información adicional estimar la cantidad total gastada en reparación de sierras para estas industrias y su error de muestreo.

6.- Para estudiar la cuantía de los impuestos pagados por un municipio, se divide éste en 150 conglomerados de tamaño 100. Se seleccionan con idéntica probabilidad y con reemplazamiento 4 de ellos, resultando:

$\sum X_{ij} = 1\,458\,236,\ 2\,351\,157,\ 1\,987\,112$ y $2\,335\,789$.

a) Estimar el total de impuestos pagados por el municipio.
b) Estimar la media de impuestos pagados por cada ciudadano.

7.- Se desea realizar un estudio de satisfacción de los clientes de una fábrica que sirve a establecimientos de 8 provincias españolas. Para ello se seleccionan con reposición y probabilidades iguales tres de éstas provincias preguntando en una escala de 0 a 10 el grado de satisfacción con sus servicios a todos los clientes de cada provincia.

	Provincia 1	Provincia 2	Provincia 3
$\sum_{i=1}^{M} X_i$	1,254	1,126	1,345
$\sum_{i=1}^{M} X_i^2$	21,723	19,897	24,657
M	97	105	90

a) Estimar el grado de satisfacción medio con la empresa.
b) Obtener una estimación de la varianza del estimador de la media

BIBLIOGRAFÍA.

El protocolo de investigación.
Lineamientos para su elaboración y análisis.
Ignacio Méndez Ramírez.
De la Namihira Guerrero.
Laura Moreno Altamirano.
Cristina Sosa de Martínez.
Editorial Trillas.
México 4ta. Reimpresión.

Como hacer una tesis.
Huáscar Taborga.
Tratados y manuales Grijalbo.
Editorial Grijalbo.
México, D.F.
Décima primera edición.
218 paginas.

Muestreo de encuestas.
Leslie Kish.
Editorial Trillas.
México.
Traducción Ricardo Vinos Cruz López.
Revisión Técnica: José Nieto de Pascual.
Tercera reimpresión (junio 1982).
736 paginas.

Some theory of sampling.
William Edwards Deming.
Dover Publications, Inc.
New York.
602 paginas.

Técnicas de Muestreo.
William Cochran Gemmell (1909-1980).
México Cecsa; 1980.
513 paginas.

Introducción al muestreo en poblaciones finitas.
Ana I. Cid Cid.
Carlos A. Delgado Manríquez.
Santiago Leguey Galán.
Editorial Ene Nuevas Estructuras,S.L.
Madrid.
276 páginas.

Técnicas de muestreo estadístico
Teoría, práctica y aplicaciones informáticas.
César Pérez López.
Alfaomega Grupo Editor.
Madrid, España.
603 páginas.