UNIVERSIDAD DE DEUSTO

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

TESIS DOCTORAL

UN ESTUDIO SOBRE LA SELECCIÓN EN EL DISEÑO DE ENCUESTAS POR MUESTREO

GONZALO SÁNCHEZ-CRESPO BENITEZ
DONOSTIA - SAN SEBASTIÁN, 1998

UNIVERSIDAD DE DEUSTO

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

TERCER CICLO

Programa: ECONOMÍA Y DIRECCIÓN DE EMPRESAS

UN ESTUDIO SOBRE LA SELECCIÓN EN EL DISEÑO DE ENCUESTAS POR MUESTREO

Tesis doctoral presentada por GONZALO SÁNCHEZ-CRESPO BENITEZ

Dirigida por el DR. D. IÑAKI GARCÍA ARRIZABALAGA

Dirigida por el DR. D. JOSÉ LUIS SÁNCHEZ-CRESPO RODRÍGUEZ

Los Directores

El Doctorando

Donostia - San Sebastián, septiembre de 1998

Dedicado a mis hijas Lara y Paula

AGRADECIMIENTOS

En un trabajo de este tipo la lista de agradecimientos debe ser necesariamente larga, pues largo es el periodo de gestación y grandes las ayudas solicitadas y recibidas.

En primer lugar quiero agradecer a mi padre y también Catedrático de Estadística, José Luis Sánchez-Crespo Rodríguez, su paciencia en la corrección y redireccionamiento constante de este trabajo, su amabilidad en las horas bajas y su elevado nivel de exigencia en los momentos pujantes. Dentro de este grupo directo de agradecimientos tengo el deber de agradecer su comprensión, aliento y buen tino en sus apreciaciones al también profesor doctor Iñaki García Arrizabalaga.

Han existido otras personas a las que he molestado con consultas y peticiones de lectura de distintas partes de este documento. En este terreno la lista es larga en nombres, si bien institucionalmente se concentran en el área de diseño de encuestas del Instituto Nacional de Estadística. Estas personas han sido: Gerardo Prieto, Javier De Parada, Juana Porras, Gonzalo de Parada,

Carlos Ballano y Montserrat Herrero. De todos ellos he recibido ideas y sugerencias que me han abierto caminos en la investigación.

Existen dos personas sin las que este trabajo no hubiese podido tener toda su virtualidad. Me han ayudado en aspectos de la programación informática Alberto Lezkano y Santiago Jiménez.

De la atenta lectura de originales quiero dejar constancia de las aportaciones realizadas por Carmen Diz. A quien agradezco especialmente sus inteligentes sugerencias.

Del trabajo de recopilación de artículos y libros quiero agradecer a la Universidad de Deusto el trabajo extraordinario de recopilar una documentación que en muchas ocasiones no estaba directamente disponible. En especial quiero agradecer sus desvelos a Carmen Navarrete por su ayuda y eficaz gestión.

También quiero agradecer los trabajos mecanográficos de Yolanda Gómez por su especial esmero en la presentación.

No obstante la ayuda recibida de todas estas personas, si existe en este trabajo alguien realmente necesario y motivador son mi mujer y mis hijas.

MOTIVACIÓN Y ESTRUCTURA

MOTIVACIÓN

Trabajo en el Instituto Nacional de Estadística. Durante dos años desarrollé mi labor en la sección de muestreo y, posteriormente, he ocupado el puesto de Delegado Provincial.

El embrión de este trabajo fue la preocupación por la forma de seleccionar las muestras de viviendas en la Encuesta de Población Activa. Se puede observar cómo la selección de las viviendas que forman parte de la muestra se lleva a cabo , en la segunda etapa, por muestreo sistemático sobre un marco ordenado por nombre de la calle. Esto en cuanto a las variables de interés, que en este caso es la población que habita en cada una de las viviendas, es equivalente a utilizar un orden aleatorio en la población. En principio este procedimiento tiene ventajas puesto que asegura un método insesgado en cuanto a selección, y asegura en la muestra una proporción representativa de hogares de una persona, de dos , etcétera... No obstante, los inconvenientes surgen al observar que existiendo falta de respuesta, debido a ausencias y negativas, las unidades suplentes se seleccionan también aleatoriamente. Este proceso lleva a sobrerepresentar viviendas con más habitantes puesto que las negativas y las ausencias se concentran en aquellas viviendas

con menos miembros. El resultado de esta forma de seleccionar las muestras puede llevar a estimar incorrectamente ciertos valores de la población, debido a que en los hogares con más de dos miembros existe una mayor tasa de paro que en hogares compuestos por una o dos personas. En este campo existen diversos estudios realizados en la sección de muestreo del INE.

Otro motivo, y tal vez el mas visible, para continuar este trabajo fue atender a personas que desde diversos ámbitos solicitaban una ayuda para diseñar la muestra que necesitaban emplear. El tipo de investigaciones solía ser de muy diferentes características, pero tenían siempre en común la preocupación por determinar las unidades últimas de muestreo. Para estas personas el procedimiento a emplear era un tema de segundo nivel, no suponía una prioridad.

El diseño de encuestas por muestreo es una técnica difícil que requiere la existencia de una infraestructura importante en cuanto a medios humanos y materiales. Dejando a un lado todo el complejo trabajo de la definición de los objetivos, de la elaboración del cuestionario, del trabajo de campo, de los manuales de entrevistadores, de los programas de entrada de datos y de tabulación, y concentrándonos únicamente en el diseño de la muestra, encontramos también la necesidad de disponer de un conocimiento experto; normalmente una persona o un equipo que conoce los

procedimientos descritos en los manuales de la teoría y práctica del muestreo en poblaciones finitas y que es capaz de recomendar una estrategia para la selección de las unidades muestrales y de formular matemáticamente el proceso de estimación. Esta persona o equipo precisa a su vez disponer de toda la información existente relacionada con la encuesta. Esto, en mucha ocasiones, requiere la elaboración de programas de ordenador a medida para evaluar la información auxiliar que permitirá mejorar el proceso de muestreo y tomar la mejor decisión en cuanto al diseño.

Con esta perspectiva cuando alguien pregunta ¿qué unidades debo investigar?, la respuesta rápida del experto no puede ser otra que recomendarle una extracción aleatoria de unidades y la aplicación de las fórmulas disponibles sobre muestreo aleatorio simple. Esto, muchas veces, es como utilizar un petrolero para llevar un saco de naranjas. Exigirá muchas más unidades muestrales que las que serían necesarias en el caso de utilizar la adecuada estrategia. Pero lamentablemente, las personas que suelen necesitar basar sus conclusiones en una muestra no suelen disponer de la infraestructura necesaria para elegir la mejor alternativa. Este trabajo pretende precisamente esto, proporcionar la infraestructura necesaria para realizar una selección muestral.

Al menos en parte esperamos haberlo conseguido.

ESTRUCTURA

Este trabajo se ha ido gestando en diferentes fases. En su primer momento se trató de un informe al Servicio de Diseño de Muestras del INE. Este informe tenia cuatro partes y una extensión pequeña. En el transcurso de la investigación se ha ido ampliando con la incorporación de diferentes aspectos que complementaban aquel estudio inicial. En la actualidad el trabajo esta dividido en quince capítulos, si bien sigue existiendo la misma estructura original. Siguiendo aquella estructura podemos distinguir cuatro grandes bloques: una «introducción», donde se establecen las posibilidades de cálculo de una aplicación para la simulación de procesos de muestreo que fue realizada, inicialmente, para servir de comprobación de la tesis principal del informe; un segundo apartado donde se explican los distintos «tipos de muestreo» que se van a considerar en el análisis; posteriormente se proporcionan los resultados obtenidos sobre distintas poblaciones de «datos»; por último, se lleva a cabo una extensión de los tipos de muestreo propuestos al «muestreo bietápico», y se analiza el efecto de los «cambios en la especificación» de los modelos.

ÍNDICE GENERAL

Pág.

Pág.

Pág.

Pág.

Pág.

Pág.

Pág.

Pág.

Pág.

ÍNDICE DE TABLAS

INDICE DE TABLAS

Pág.

Pág.

Pág.

Pág.

ÍNDICE DE GRÁFICOS

INDICE DE GRÁFICOS

Pág.

Pág.

Pág.

CAPITULO 1

INTRODUCCIÓN GENERAL

1.1. OBJETIVOS DE LA INVESTIGACIÓN

El objetivo principal de esta investigación consiste en evaluar diferentes métodos de selección sistemáticos de unidades muestrales. Mediante la utilización de un programa de ordenador capaz de incorporar la información existente sobre la estructura de la población objetivo en un análisis de superpoblación.

Podemos desglosar este objetivo en cinco grandes bloques, en función del problema o de la dificultad con la que nos hemos encontrado:

En primer lugar, se puede observar cómo en la práctica de las encuestas por muestreo se aplica con frecuencia la técnica del muestreo sistemático. Esto es debido a las ventajas operacionales que presenta este método y a ciertas propiedades que verifica bajo determinadas supuestos. Sus principales ventajas son: su facilidad de aplicación; recoge un posible efecto de estratificación; extiende la muestra a toda la población; si las unidades de muestreo se ordenan conforme a una variable conocida y relacionada estructuralmente con

la variable de estudio, se pueden obtener grandes ganancias en precisión. Entre sus inconvenientes se citan la posibilidad de pérdidas en precisión debidas a periodicidades ocultas, siendo especialmente sugeridas precauciones en el caso de la existencia de ciclos en la población. Este tipo de poblaciones son más frecuentes de lo que se piensa en la práctica de las encuestas por muestreo. Un ejemplo de esta situación es cuando el marco de unidades es producto de una operación censal y dentro de los hogares figuran sus miembros clasificados en padres e hijos por orden de edad (Murthy, 1988:157). Si los grupos son de un tamaño similar y el intervalo de muestreo coincide con esta periodicidad el muestreo sistemático sería muy ineficaz. En este trabajo hemos planteado un método que elimina el efecto tendencia, que estaba ya en la literatura sobre muestreo (ver correcciones de Yates, método centrado de Madow, muestreo sistemático equilibrado o muestreo modificado) y que, además, tiende a eliminar también el efecto cíclico, de forma que para corregirlo no es necesario conocer a priori su existencia o sus características. También hay que resaltar que el método propuesto presenta unos resultados robustos en cuanto a cambios en las especificaciones del modelo.

La segunda dificultad con la que nos enfrentamos fue, la situación que se presenta en la teoría del muestreo en poblaciones finitas, al estimar el error debido al muestreo cuando el esquema de selección de la muestra es sistemático. Este es el caso en el que sólo se dispone de una unidad por estrato o de un único conglomerado en la muestra. Una ilustración de que estimar la varianza en esta situación no es posible, en un sentido estricto (Cochran:1977, 278). Sin embargo la teoría clásica considera distintas soluciones para aproximar una estimación acurada de la varianza del muestreo sistemático, si bien en todas ellas se precisa realizar algún tipo de supuesto sobre la estructura de la población.

Tenemos así el tercer bloque, o la tercera dificultad que es el disponer de un instrumento que permita evaluar la población marco para estudiar si se verifican las hipótesis que se realizan sobre su estructura. Es necesario tener en cuenta que no sólo depende del tamaño de las muestras sino también del número de clases que se formen, y que los cálculos que se precisan no están disponibles mediante los programas estadísticos de uso general.

La cuarta dificultad consiste en, una vez establecidos los supuestos sobre la estructura de la población y evaluadas las alternativas sobre tamaño de muestra y número de clases, evaluar los distintos métodos de selección que se consideren como alternativos. A este efecto es necesario considerar la población dentro de un enfoque de modelo de superpoblación y determinar la formulación estadística que mejor se ajuste a esa determinada estructura poblacional que se esté considerando.

La última dificultad con la que se enfrenta este trabajo considera la extensión al muestreo bietápico, haciendo referencia al esquema general de selección de muestras en el diseño de muchas de las principales encuestas que se llevan a cabo en la práctica. Esto es, la selección de las unidades de primera etapa se realiza con probabilidades proporcionales al tamaño y las unidades de segunda etapa con un esquema sistemático. Para ello y para completar el diseño de selección de unidades muestrales se incorporan algunos métodos con probabilidades proporcionales al tamaño.

1.2. CONCEPTOS UTILIZADOS EN ESTE TRABAJO

Por «población» entenderemos el conjunto de unidades sobre el que se desea obtener información. Supondremos que se dispone de una lista con información de cada unidad poblacional procedente de un censo, registro administrativo o investigación anterior a la selección de la muestra. Esta población se define como «población objetivo», y no se limita a poblaciones naturales o a poblaciones generadas artificialmente con procedimientos aleatorios, sino que también pueden ser poblaciones generadas de forma que representan cierta estructura poblacional. Esta población objetivo, puede ser en determinados casos considerada como la realización concreta de una muestra.

Utilizaremos la expresión «estructura poblacional», dentro del ámbito de los modelos de superpoblación, para referirnos a los parámetros que definen una población natural. De esta manera será posible obtener, con la definición de ciertos parámetros, simulaciones aleatorias de la población original.

Cada unidad que compone la población está caracterizada por uno o varios valores. Estas características, con los valores que toma, se denominarán «variables». En este estudio se considerarán variables cuantitativas y cualitativas. Las variables se pueden clasificar según su utilización en el esquema de selección como variables de estudio o variables auxiliares. La variable elegida como de estudio será aquella de la que se desea información actualizada. La variable auxiliar será aquella que por su especial relación con la variable objeto de estudio y por su carácter estructural permite auxiliar el proceso de selección de una muestra. Esta última variable se utilizará bien para calcular los valores de las probabilidades de selección de cada unidad poblacional, o bien para ordenar las citadas unidades poblacionales.

El espacio formado por todas las muestras que es posible obtener con un determinado procedimiento de selección constituye el «espacio muestral». El «espacio paramétrico» lo consideraremos formado por todos los parámetros que intervienen en el proceso de estimación.

Por «muestra» entenderemos una parte representativa de la población elegida con un determinado procedimiento de selección.

En cuanto a los «procedimientos de selección» los entenderemos como modelos de selección, deberán permitir conocer la probabilidad de una determinada muestra y formar el espacio de todas las muestras posibles dado el procedimiento. Por tanto se refiere al conjunto de especificaciones que deben ser tenidas en cuenta a la hora de seleccionar una muestra y de construir los estimadores.

En este estudio también utilizaremos repeticiones del proceso de selección de una muestra. Denominaremos a estas repeticiones como «reiteraciones». Esto es, con esta expresión se identifica el número de veces que se va a repetir el proceso de selección de muestras. Evidentemente en el trabajo de campo de las encuestas por muestreo, únicamente se investiga una muestra. Sin embargo, en los trabajos asociados al diseño de una encuesta, si se dispone de información de la estructura poblacional, es posible repetir el proceso de selección tantas veces como se desee para poder analizar como varía el estimador. Con este proceso de simulación se trata de observar y cuantificar como varía el estimador y sus parámetros asociados, no sólo con los datos de una sola muestra, sino con la variación real producida al variar la muestra seleccionada.

En ciertos casos, al no poder generar todo el espacio muestral utilizaremos una «representación del espacio muestral». Esta consistirá en una selección aleatoria del conjunto de todas las muestras posibles. En los casos de muestreo aleatorio, la formación completa del espacio muestral puede llegar a ser excesivamente compleja. Así, en estos casos, muestreo sin o con reposición, estratificado con una unidad por estrato o muestreo bietápico, en la aplicación POSDEM se ha optado por trabajar con una representación del espacio muestral

El error cometido al sustituir la información que proporciona toda la población por lo que proporciona una única muestra se denominará «error de muestreo». Tiene asociados los conceptos de «precisión» y «acuracidad». En el caso de utilizar estimadores insesgados ambos conceptos coinciden y se pueden medir por el error cuadrático medio, que coincidirá con la varianza del estimador. En caso contrario para medir el error tendremos que sumar a la varianza el sesgo para obtener el error cuadrático medio. El error de muestreo se define como la raíz cuadrada de la error cuadrático medio del estimador.

Un «estimador insesgado» lo es cuando su esperanza matemática coincide con el valor del parámetro poblacional. A efectos de

esta aplicación un estimador será insesgado cuando al repetirse el proceso de muestreo, mediante reiteraciones, un número suficiente de veces la diferencia entre el valor medio de estas reiteraciones y el valor objetivo poblacional tienda a cero. En el caso de muestreo sistemático, donde las reiteraciones permiten obtener todo el espacio muestral, la diferencia, cuando el estimador es insesgado, será cero, y por tanto, la media del estimador coincidirá con el parámetro poblacional que se quiere estimar.

Con el «error cuadrático medio de un estimador» representaremos la medida del error, en términos de probabilidad, que se comete al sustituir un valor poblacional, desconocido, por una estimación basada en los datos de una muestra. Su raíz cuadrada es el error de muestreo. En términos relativos se utiliza el coeficiente de variación, que al ser independiente de la unidad de medida permite comparaciones entre distribuciones diferentes. Al repetir el proceso de muestreo cada muestra proporciona un valor distinto del estimador y de su error de muestreo. Como cada muestra tiene una determinada probabilidad de ser seleccionada, el valor del estimador asociado con dicha muestra tendrá esa misma probabilidad de ocurrir. Por tanto, los valores posibles del estimador con sus correspondientes probabilidades

forman una variable aleatoria, y se podrá calcular su varianza, desviación típica ... etc. Se constituye así la distribución del estimador en el muestreo.

Utilizaremos «intervalos de confianza» del 95 % definidos como intervalos aleatorios de forma que de cada cien muestras obtenidas con el mismo procedimiento se cubrirá en noventa y cinco muestras el valor del parámetro que queremos estimar. Así, se define el intervalo de forma que el valor real del parámetro a estimar esté comprendido dentro de dicho intervalo, con una probabilidad igual al noventa y cinco por ciento.

Como último concepto dentro de este apartado queremos destacar el concepto de «programa de ordenador POSDEM[1]» se trata de un conjunto de instrucciones redactadas en un lenguaje de programación, visualbasic, que permite al usuario seleccionar alternativas y obtener resultados en un campo concreto, en este caso en el diseño de encuestas por muestreo. Lleva a cabo tareas que por

[1] Programa para Optimizar la Selección en el Diseño de Encuestas por Muestreo . En su versión Windows este programa se ha desarrollado en colaboración entre Alberto Lezkano y Gonzalo Sánchez-Crespo

su repetición, complejidad de calculo o dificultad no podrían llevarse a cabo de otro modo. Es un instrumento de productividad personal.

1.3. PROBABILIDADES DE SELECCIÓN: IGUALES O DESIGUALES

Cada unidad de la población debe tener una probabilidad conocida de selección para que el muestreo sea probabilístico y puedan estimarse los errores debidos al mismo. En el caso de probabilidades iguales todas las unidades de la población tienen la misma probabilidad de ser elegidas para formar parte de una muestra. Estos métodos tienen la ventaja de mantener la probabilidad de selección por lo que la estimación, tanto de los parámetros que se quieren estimar como de sus errores de muestreo, es relativamente sencilla, a excepción de los diseños sistemáticos. Por contra, existen métodos de selección donde las probabilidades son desiguales. Esto es, cada unidad tiene una probabilidad distinta de pertenecer a la muestra. Estos métodos tienen la ventaja de incorporar información adicional y, generalmente, disminuir el error debido al muestreo.

1.4. MÉTODOS DE SELECCIÓN DE UNIDADES

Los principales esquemas de selección de muestras pueden clasificarse en función de las probabilidades de selección asignada a cada unidad poblacional y del método utilizado para la selección. En este apartado realizaremos una breve descripción de los métodos que serán descritos más formalmente en sus capítulos correspondientes. Así podemos empezar describiendo los métodos cuando las probabilidades de selección son iguales:

1.- «Muestreo aleatorio simple»: en distintos libros se dan definiciones diferentes de muestreo aleatorio simple. En este apartado consideraremos, dentro de este epígrafe, el muestreo con y sin reposición y probabilidades iguales. Estos métodos se pueden definir en función de si las unidades seleccionadas se devuelven a la población, en este caso tendremos un esquema con reposición; o, en el caso de que las unidades, una vez analizadas, no se reincorporen a la población, y por tanto su probabilidad de selección para ocasiones sucesivas sea cero una vez seleccionada, en este segundo caso tendremos un esquema de selección que se denomina sin reposición.

Con el método con reposición una unidad puede pertenecer a la muestra más de una vez. Coincide en sus fórmulas con el muestreo sin reposición cuando la población es infinita. El muestreo sin reposición es más preciso que el método anterior debido al factor de corrección que se aplica a sus fórmulas para estimar el error de muestreo. En ambos casos cada unidad poblacional tiene una probabilidad conocida e igual de selección.

2.- «Muestreo sistemático»: este método es muy conveniente cuando las unidades de la población se encuentran numeradas serialmente de acuerdo con una variable relacionada con las variables de estudio. En ese caso puede presentar importantes ganancias en precisión. Bajo la denominación de muestreo sistemático se encuadran diferentes esquemas, algunos muy diferentes entre sí. Esto dificulta proporcionar una definición general. No obstante, en este trabajo consideraremos como sistemáticos básicos aquellos métodos caracterizados por la selección aleatoria de una primera unidad poblacional, las restantes se seleccionan conforme un determinado procedimiento matemático establecido previamente.

Siguiendo a Murthy (1988)[2], el muestreo sistemático, en general, es una técnica de muestreo más conveniente que el muestreo aleatorio simple y, al mismo tiempo, asegura para cada unidad igual probabilidad de inclusión en cada muestra. La conveniencia de este método viene dada por la simplicidad de obtención de la muestra. Esta es una conveniencia operacional de considerable importancia en el trabajo de campo de las encuestas a gran escala por muestreo. El método sistemático presenta también la ventaja de aportar estimadores más eficientes que los proporcionados por el muestreo sin reposición bajo ciertas condiciones que se dan habitualmente en la práctica.

Dentro de este esquema general consideraremos los métodos de intervalo de muestreo constante, intervalo variable, sistemático equilibrado, sistemático modificado, muestras centradas de Madow, y el método de las correcciones finales de Yates.

Consideraremos los siguientes tipos de muestreo sistemático:

[2] (Krishnaiah, Rao, 1988)

2.1.- «Muestreo sistemático con intervalo de selección constante»: es el método tradicional. Divide la población en grupos. Si el tamaño de población se representa con N y el tamaño de muestra con n, entonces el intervalo de selección se define con la expresión: $k = N/n$. Se selecciona una unidad del primero de los grupos, después se seleccionan el resto de las unidades conforme la posición que ocupan respecto de esta primera unidad, sumándole un período constante (k). Formalmente: la selección se lleva a cabo obteniendo un número aleatorio entre 1 y k, que denominaremos i. Este número permite determinar la primera unidad que figurará en la muestra. Las n-1 unidades restantes se seleccionan de forma que:

$$zc = i + (j-1)\,k$$

Donde

zc= Valor que identifica las unidades seleccionadas con intervalo constante

i= número aleatorio de selección. $1 \leq i \leq k$

j= número correlativo entre 1 y n. Donde n es el tamaño de muestra.

k= Tamaño de los grupos formados para la selección. k= N/ n

2.2.- «Muestreo sistemático con intervalo de selección variable»: es una variante del método anterior. En lugar de sumar un período constante se suma un período variable, conforme cierta regla. Este método tiene las siguientes ventajas: mantiene las probabilidades de selección; presenta un óptimo, en poblaciones del tipo $X_i=i$; $\forall$ i=1,2,3 ... N, puesto que para un tamaño de muestra igual a la raíz cuadrada de N, cualquiera de las muestras posibles proporciona el verdadero valor del parámetro. Con una notación similar a la empleada anteriormente:

$$zv = i + (j-1)(k+1) - c\,k$$

zv= Valor de la unidad muestral seleccionada con intervalo variable

Donde, por definición, los valores que toma c vienen dados por las siguientes situaciones:

Si zv=jk no ha sucedido nunca, entonces c=0

Si zv=jk ha ocurrido una vez, entonces c=1

Si zv=jk ha ocurrido dos veces, entonces c=2 ....

Así, c= 0,1,2 ... de acuerdo con el número de veces que ha ocurrido que zv=jk

2.3.- «Muestreo sistemático equilibrado»: siguiendo a Murthy (1967:165)

La muestra sistemática i-ésima consistirá en las unidades

$i+2jk,\ 2(j+1)k-i+1$ $(j=0,...(1/2)n-1/2)$ para n par y

$i+2jk,\ 2(j+1)k-i+1,\ i+(n-1)k$ $(j=0,...(1/2)n-3/2)$ para n impar

2.4 «Muestreo sistemático modificado»: en este caso la muestra sistemática i-ésima consistirá en las unidades

$i+jk,\ N-i-jk+1$ $(j=0,...(1/2)n-1/2)$ para n par y

$i+jk, N-i-jk+1,\ i+(1/2)(n-1)k$ $(j=0,...(1/2)n-3/2)$ para n impar

2.5.- «Muestreo sistemático centrado»: consiste en, cuando el número de muestras que es posible formar es par, seleccionar como espacio muestral completo las dos muestras centrales. Así, si el número de grupos que es posible formar es k, seleccionamos las muestras que ocupan las posiciones k/2 y k/2+1 de forma que cuando se desea obtener una muestra con este procedimiento se selecciona una de estas dos con probabilidad 1/2. En el caso de k impar se selecciona la que ocupa la posición central. A efectos de calculo computerizado de la

aplicación POSDEM se ha optado por formar el espacio muestral con las tres muestras que ocupan el lugar central para el caso k impar y con las dos muestras centrales para el caso de k par.

2.6.- «Muestreo sistemático corregido»: este es el caso que se conoce por las correcciones de Yates. Consiste en obtener el espacio muestral conforme al procedimiento de muestreo sistemático con intervalo constante y corregir el estimador ponderando el primero y el último miembro por:

$$1 \pm \frac{n(2i-k-1)}{2(n-1)k}$$; i es el aleatorio entre 1 y k.

El signo se reserva: + para el primer miembro y - para el último.

2.7.- «Muestreo sistemático circular»: es un instrumento utilizado para superar la dificultad que representa que el tamaño de la población, N, no sea múltiplo del tamaño de muestra, n. El procedimiento consiste en elegir aleatoriamente un arranque, r, entre 1 y N, la muestra consiste en las unidades correspondientes a los números

$r+jk$ si $r+jk \leq N$

y con

$r+jk-N$ si $r+jk>N$

para $j=0,1,2, \ldots(n-1)$

Este procedimiento fue propuesto por D.B. Lahiri (1952). El intervalo k se determina habitualmente como el entero más próximo a N/n para asegurar una extensión adecuada de la muestra sobre todo el marco muestral.

Siguiendo a Azorín (1969: 219), vemos que cuando el tamaño de la población no sea múltiplo del de la muestra y se verifique, por ejemplo, $N=nk+r$, la media ya no será un estimador insesgado de la media poblacional. Para evitar el sesgo habría que dar a r muestras la probabilidad $(n+1)/N$ y n/N a las $k - r$ muestras restantes.

Por simplicidad asumiremos que N es múltiplo de n. No obstante los resultados obtenidos pueden generalizarse al caso donde N no sea múltiplo de n. La aplicación POSDEM utiliza por defecto la opción N múltiplo, pero es posible cambiar la configuración para conseguir muestras sistemáticas circulares con los procedimientos sistemáticos descritos.

En la aplicación POSDEM se ha programado la posibilidad de aplicar el procedimiento circular, además de para el caso del muestreo

sistemático con intervalo constante, para los casos de muestreo sistemático con intervalo variable, equilibrado y modificado.

3.- «Muestreo de conglomerados»: básicamente, con este método se realiza la selección de unidades poblacionales siendo éstas a su vez grupos de elementos. En este trabajo se ha incorporado este método como auxiliar del muestreo sistemático. Esto es, se ha considerado que las unidades poblacionales estaban compuestas por conglomerados formados con arreglo a los cuatro métodos sistemáticos descritos en el apartado anterior. A efectos de simplicidad y dado que supone un esquema general con múltiples ventajas, se ha optado por mantener constante e igual a dos el número de conglomerados en la muestra. Esto, sin embargo, no supone una limitación en cuanto al número final de unidades primarias en la muestra que dependerá del número de unidades en cada conglomerado. También se han considerado todos los conglomerados de igual tamaño.

Así, se consideran modelos de muestreo de conglomerados sistemático con intervalo constante o variable, equilibrado o modificado. Cada muestra sistemática puede considerarse como un conglomerado. Para realizar una selección con el método de conglomerados se ha

recurrido a un procedimiento que consiste en seleccionar una muestra compuesta por dos semimuestras sistemáticas. La primera muestra de dos conglomerados estará formada de la siguiente forma: la primera mitad de la muestra se corresponderá con la primera de las semimuestras sistemáticas obtenidas con tamaño de muestra n/2 y la segunda parte de la muestra de dos conglomerados estará formada por la primera semimuestra de la segunda mitad de las semimuestras sistemáticas obtenidas, el resto, correlativamente. El procedimiento por el que se forman las muestras con dos conglomerados se describe con un ejemplo sencillo en su apartado correspondiente. No obstante la formulación matemática que se utiliza es la indicada en cada método sistemático descritos anteriormente.

4.- «Muestreo estratificado con una unidad por estrato»: es especialmente interesante la comparación del muestreo sistemático con el muestreo estratificado aleatorio con una unidad por estrato. Para obtener el error de muestreo del estimador media, para el caso en el que la selección de las unidades que forman la muestra se obtengan mediante un procedimiento de muestreo estratificado con una unidad por estrato, se puede utilizar la siguiente expresión, (Cochran, 1977: 262):

$$V(\bar{\bar{x}}_{str}) = (\frac{N-n}{N})\frac{S^2_{wst}}{n}$$

Donde

$$S^2_{wst} = \frac{1}{n(k-1)} \sum_{h=1}^{n} \sum_{i=1}^{k} (x_{ih} - \bar{x}_{.h})^2$$

con h representando el estrato h-ésimo.

5.- «Muestreo con probabilidades desiguales»: es óptimo cuando las probabilidades de selección son exactamente proporcionales a los valores de la variable que se quiere estimar. En ese supuesto, sea cual sea la muestra elegida, el error de muestreo es cero. Por tanto si la variable que se desea utilizar para asignar probabilidades de selección a cada unidad poblacional está relacionada con el valor de la variable que se quiere estimar, el método mejorará la precisión del estimador. Normalmente se utilizan probabilidades proporcionales al tamaño, ya que esta suele estar correlacionada con la variable de estudio.

6.- «Muestreo bietápico»: consiste en seleccionar una muestra de unidades de primera etapa, y, posteriormente, realizar dentro de estas unidades un nuevo proceso de muestreo para seleccionar las unidades últimas. En este caso la selección de la muestra se lleva a cabo

en dos etapas. Cada una de las cuales incorpora un cierto grado de variabilidad a la estimación.

1.5. POBLACIONES ORDENADAS: TENDENCIA LINEAL, POBLACIONES CORRELACIONADAS Y POBLACIONES CON VARIACIÓN PERIÓDICA

La ganancia en precisión obtenida al ordenar las unidades poblacionales, en base a una variable auxiliar, es debida a la presencia de una tendencia en los valores de las unidades. Se puede seguir el ejemplo con la población de las aldeas (Murthy,1988:161), que permite observar la nube de puntos para la variable "área cultivada" ordenados por "áreas geográficas" en el que existe tendencia con heterocedastícidad y para la variable "población 1961" ordenados por "población en 1951" donde se observa tendencia exponencial.

Consideraremos, a continuación, algunos casos particulares importantes:

1.- «Tendencia lineal»: en este caso Cochran(1977) demuestra un teorema según el cual la mayor precisión corresponde al muestreo estratificado aleatorio cuando existe tendencia lineal, en comparación con los métodos sin reposición y sistemático.

Posteriormente se han propuesto métodos como el centrado de Madow, de las correcciones extremas de Yates o los métodos sistemáticos equilibrado y modificado que permiten corregir el efecto tendencia en el muestreo sistemático, si bien bajo determinados supuestos.

2.- «Poblaciones correlacionadas»: con este coeficiente se trata de expresar la mayor semejanza que existe entre dos observaciones cuando están más próximas que cuando están más alejadas. Los coeficientes de correlación intraclásica representan el cociente de la covarianza relativa a la distribución conjunta de la variable aleatoria correspondiente a la observación y de la de otra cuya distancia a la primera sea d, por la varianza común a ambas variables. La representación de p_d en función de d recibe el nombre de correlograma. Siguiendo a Azorin(1969) diremos que la población está autocorrelacionada cuando se verifique $p_d<p_e$ siempre que sea $d<e$. En este tipo de poblaciones el muestreo estratificado aleatorio es superior al aleatorio simple, pero no pueden establecerse consecuencias generales en lo que se refiere al muestreo sistemático.

Cochran(1946), obtuvo resultados numéricos para la precisión relativa del muestreo estratificado y sistemático en

poblaciones con correlogramas lineales y exponenciales. El teorema de Cochran (1946) generalizado por Quenouille (1949) demuestra que el muestreo sistemático da estimaciones más precisas que el estratificado aleatorio si el correlograma es cóncavo hacia arriba. Una de las funciones más simples de este tipo es $p_d=e^{-d}$ empleada por Osborne (1942). Otra, también sencilla, es :

$$p_d = \frac{a_1 + a_2 d}{a_3 + a_4 d}$$

uno de cuyos casos particulares es la recta:

$$p_d = 1 - \frac{d}{a}$$

3.- «Poblaciones con variación cíclicas»: en cuanto al caso de poblaciones con variación cíclica, siguiendo a Murthy(1988:169), si en una población las unidades se siguen unas a otras conforme un patrón regular y repetitivo, ciclo, entonces el muestreo sistemático debe utilizarse con considerable cuidado.

Si la periodicidad de la curva es conocida y la amplitud del intervalo es múltiplo impar de la mitad del periodo, la varianza se reduce considerablemente.

Ejemplos de este tipo de población se obtienen cuando ésta consiste en grupos de igual o aproximadamente igual número de unidades y las unidades dentro de cada grupo se ordenan de acuerdo con un patrón. Así, las poblaciones que provienen de enumeraciones censales, donde el universo de hogares esta compuesto por personas normalmente identificadas por orden de padre, madre, hijos de mayor a menor. Cuando el intervalo de muestreo coincide con el tamaño del grupo familiar, la muestra tendrá una composición similar y el muestreo sistemático tendrá un comportamiento ineficaz. En este caso la eficiencia relativa del muestreo sistemático sería mínima en comparación con el muestreo aleatorio simple. El caso más favorable se presentaría cuando el intervalo de muestreo k fuese un múltiplo impar del semiperíodo.

Poblaciones con periodicidad, más o menos regular, no son extrañas en la práctica de las encuestas por muestreo. Es necesario estudiar la población con mucho cuidado antes de usar muestreo sistemático. Este tipo de riesgo ha sido puesto de manifiesto, entre otros, por Sthephan, Deming y Hansen(1940) y Lahiri(1954A).

La situación en la cual el muestreo sistemático, en relación al muestreo aleatorio simple, es probablemente más eficiente, además

de ser operacionalmente más sencillo, es aquella en la cual la nube de puntos de los valores de una variable asociada presentan una tendencia o muestran oscilaciones no muy pronunciadas alrededor de una tendencia lineal o curvilínea, y donde el intervalo de muestreo utilizado sea menor que el período más pequeño observado en las oscilaciones.

Cochran(1977:217-219), explica distintos ejemplos de poblaciones periódicas: el flujo de tráfico que pasa por un punto dado; las ventas de un almacén sobre un periodo de siete días. En estos casos recomienda extender la muestra a lo largo de la curva periódica. En el ejemplo de las ventas en un almacén se deberá cumplir que cada día reciba la misma representación. Por tanto es necesario estudiar la estructura periódica que se puede presentar en la población. Este autor recomienda que si no se conoce y existe la posibilidad de variación periódica es mejor aplicar muestreo aleatorio simple o muestreo estratificado.

Otros autores, como Madow(1946), Finney(1948) o Milne(1959), observan casos menos obvios que los anteriores de «casi periodicidad» que hace que el muestreo sistemático presente

resultados mediocres en ciertos valores de tamaño de muestra y un buen comportamiento en otros.

1.6. COMPORTAMIENTO DE LA VARIANZA. POBLACIONES NATURALES

Con los datos de una sola muestra el muestreo sistemático presenta dificultades para estimar la varianza del estimador. Un procedimiento útil consiste en estimar el error uniendo dos unidades contiguas de cada grupo, esto es, considerando que disponemos de dos unidades muestrales por estrato. De esta forma uniendo dos unidades muestrales consecutivas se puede mejorar la estimación del error de muestreo, del método sistemático con intervalo constante o intervalo variable, en contraposición a cuando se utiliza la fórmula del muestreo sin reposición.

Se puede comprobar que para que el muestreo sistemático sea eficiente es necesario que el coeficiente de correlación entre pares de unidades sea elevado y negativo. Alternativamente, podemos comprobar que el muestreo sistemático será considerablemente ineficaz si el coeficiente de correlación intraclásica fuese positivo. Un método para conseguir un coeficiente de

correlación intraclásica negativo elevado consiste en ordenar las variables en orden ascendente o descendente, de acuerdo con una característica auxiliar asociada con las variables objeto de estudio. Puede observarse que el comportamiento de la varianza no es tan regular como en el caso del muestreo aleatorio sin reposición, debido a que no depende sólo del tamaño de muestra y de la varianza poblacional, sino también del coeficiente de correlación intraclásica que depende, a su vez, del tamaño de muestra y de la ordenación de las unidades. Se pueden analizar diferentes ejemplos empíricos con poblaciones naturales para ilustrar la conducta de la varianza cuando aumenta el tamaño de la muestra para diferentes ordenaciones.

Las poblaciones que hemos utilizado en nuestro trabajo en diferentes campos y con distintas finalidades han sido principalmente:

1.- «Franjas de bosque y volumen de madera». Variable de estudio: "el volumen de madera". Tamaño de población: 176 "franjas de bosque". Variable auxiliar: " el ancho de franja". En esta población se espera que el volumen de madera esté relacionado con el ancho de franja. Ordenar la población según el ancho de franja será similar a ordenar conforme al volumen de madera. Los datos se encuentran en (Murthy,1967:131) y los resultados en la página149.

Se puede observar en estos resultados que, efectivamente, el comportamiento de la varianza del muestreo sistemático es irregular, al contrario de lo que ocurre con el muestreo sin reposición. Se puede observar también como el ordenar la población ha supuesto una ayuda considerable para reducir la varianza.

2.- Aldeas y censos de 1951 y 1961. EL tamaño de la población es de 128 aldeas. Las variables del censo de 1951 son: "área demográfica en millas cuadradas", "área cultivada en acres", "núm. de personas". Las variables del censo 1961: son "número de personas", "número de cultivadores", "trabajadores en industrias familiares" y "número de hogares". Las variables de estudio son: "área cultivada" y "número de personas del censo de 1961". La población se considera con tres criterios diferentes: en primer lugar como en el marco, después en orden creciente del "área geográfica", y por último, según el "número de personas en 1951" en orden creciente. Los datos figuran en (Murthy,1967:128), y los resultados en la página 151.

Se puede observar que la ordenación de las unidades poblacionales más eficiente es el del "área geográfica" para estimar el "área cultivada", mientras que las ordenaciones según "número de

personas del año 1951" es más eficiente que los otros dos tipos de ordenación para estimar la población de 1961. Otro resultado importante consiste en observar como, cuando la ordenación es la del marco, el muestreo sistemático es menos eficaz que el muestreo sin reposición para tamaños de muestra pequeños.

3.- «Fábricas, número de trabajadores, capital y producción». Tamaño de población 80 fábricas. página 228 del mismo autor.

4.- «Censo de población de 1991 por secciones censales en la Comunidad Autónoma de Cantabria». El número de secciones es de 402. Las variables consideradas han sido las relativas a: "número de personas censadas", el número de personas en relación con la actividad: "número de activos", "de parados", "de ocupados", y parados por edad "menores de 16 años" y "mayores de 16 años". El número de secciones es de 402 correspondientes al total de la Comunidad Autónoma de Cantabria. Se ha utilizado también una población definida de igual forma para la Comunidad Autónoma del País Vasco con un tamaño de 1600 secciones censales, clasificadas según cada una de las tres provincias que la componen.

5.- «Hogares en una determinada sección censal». Variable considerada: "número de personas en el hogar". Tamaño de población: 400 hogares.

6.- «Empresas Industriales por número de trabajadores». Tamaño de población: 160 empresas. Variable de estudio "número de trabajadores".

7.- «Secciones Censales clasificadas por número de hogares en la sección». Tamaño de población: 20 secciones. Variable de estudio "número de hogares".

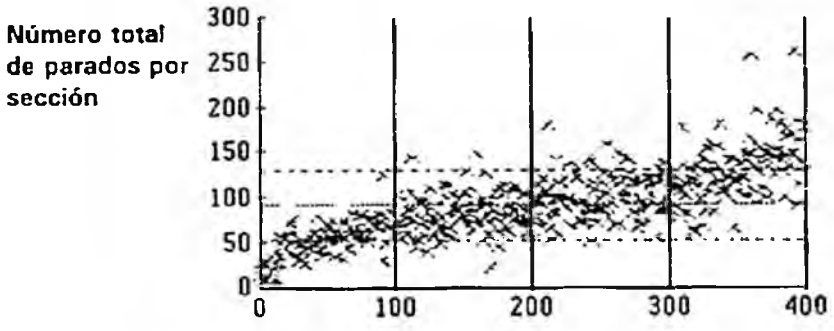
8.- «Secciones Censales clasificadas por número de personas en la sección». Tamaño de población: 16 secciones. Variable de estudio "número de parados censados en 1991" y variable auxiliar "población de la sección".

Estas poblaciones se utilizan en diferentes partes del programa para comprobar cálculos e ilustrar procedimientos.

1.6.1. VARIANZA DEL ESTIMADOR PARA UNA POBLACIÓN NATURAL

En este primer gráfico se representan todas las unidades de la población natural. Se trata de las 400 secciones censales existentes en Cantabria en el año censal de 1991, ordenadas por tamaño de población y representadas por el número de parados censados en cada una de ellas. Tendremos como parámetros básicos la media poblacional de parados por sección igual a 90.98 y la cuasivarianza poblacional igual a 1503.65

Gráfico 1.- Variable de estudio para la población natural[3]



[3] Fuente INE. Cdrom: Los municipios CERCA España Censos 89-91

En esta población de tamaño 400, con la aplicación POSDEM, es posible calcular la varianza del estimador media para distintos métodos de selección alternativos y diferentes tamaños de muestra. Concretamente para un tamaño de muestra igual a 20 unidades los resultados se presentan en la siguiente tabla.

Tabla 1.- Varianza del estimador. Valor objetivo.

MÉTODOS DE SELECCIÓN	Vp[4]	Efr sr[5]	Efr str[6]
Muestreo sin reposición (S.R.)	71.42	100	267
Muestreo estratificado con una unidad	26.75	37.5	100
Sistemático intervalo cte. (SIC)	20.93	29.3	78.2
Sistemático intervalo var. 1 (SIV1)	23.32	32.7	87.1
Sistemático intervalo var. 2 (SIV2)	23.08	32.3	86.2
Sistemático centrado (CLS)	8.24	11.5	30.8
Sistemático centrado int. var. 1 (CLSIV1)	16.62	23.2	62.1
Sistemático centrado int. var. 2 (CLSIV2)	5.19	7.2	19.4
Sistemático corrección de Yates (CY)	19.53	27.3	72.9
Sistemático equilibrado (BSS)	30.42	42.6	113.7
Sistemático modificado (MSS)	21.97	30.8	82.1
Prob. desiguales con reposición y ppt (CRPPT)[7]	29.42	41.2	110.0
Prob. desiguales sin reposición y ppt (SRPPT)	28.28	39.6	105.7
Prob. desiguales con reposición parcial y ppt CRPPPT)	28.83	40.4	107.7

4 Vp es la varianza del estimador media calculada para cada método. Es un valor objetivo poblacional.

5 Efr sr : representa la eficiencia relativa respecto al muestreo sin reposición como cociente de varianzas

6 Efr str : representa la eficiencia relativa respecto al muestreo estratificado con una unidad por estrato como cociente de varianzas

7 Por procedimientos de cálculo, al no basarse en todo el espacio muestral completo, los resultados de los métodos con probabilidades desiguales pueden presentar variaciones de unas realizaciones a otros. No obstante se ha aumentado el número de muestras obtenidas para representar el espacio muestral hasta 200 para asegurar cierta estabilidad de los valores objetivos poblacionales.

una ganancia en precisión notable. En esta primera tabla podríamos contestar, parcialmente puesto que se trata de una población natural concreta y por tanto específica, a la pregunta de cual es el error debido al muestreo que vamos a tener en esta población si en lugar de investigarla completamente analizamos únicamente una muestra.

1.6.2. VALOR ESPERADO DEL ESTIMADOR DE LA VARIANZA

Para esta población que estamos considerando es posible calcular para cada muestra el estimador de la varianza que cada método propone. Dado el espacio muestral completo podremos obtener su valor esperado. En la medida que el método de selección disponga a su vez de un método de estimación del error con los datos de una sola muestra, este valor esperado coincidirá con el obtenido como valor objetivo o valor poblacional.

Tabla 2.- Esperanza del estimador de la varianza.

MÉTODOS DE SELECCIÓN	[a]Evc	Vp[b]	Vp/Evc%
Sistemático intervalo cte. (SIC)	73.95	20.93	28.30%
Sistemático intervalo var. 1 (SIV1)	73.83	23.32	31.59%
Sistemático intervalo var. 2 (SIV2)	73.84	23.08	31.26%
Sistemático centrado (CLS)	76.39	8.24	10.78%
Sistemático centrado int. var. 1 (CLSIV1)	75.97	16.62	21.87%
Sistemático centrado int. var. 2 (CLSIV2)	71.70	5.19	7.23%
Sistemático corrección de Yates (CY)	73.95	19.53	26.41%
Sistemático equilibrado (BSS)	73.47	30.42	41.40%
Sistemático modificado (MSS)	73.90	21.97	29.73%
Prob. desiguales con reposición y ppt (CRPPT)	28.53	29.42	103.12%
Prob. desiguales sin reposición y ppt (SRPPT)	27.97	28.28	101.11%
Prob. desiguales con reposición parcial y ppt (CRPPPT)	28.30	28.83	101.87%

[a] Evc es la esperanza del estimador de la varianza. Para diseños sistemáticos se ha utilizado la técnica del lazo

[b] Vp es la varianza del estimador. Valor objetivo poblacional

Podemos observar que, para esta población, aquellos métodos que en la tercera columna presentan un valor entorno al 100% disponen de un estimador adecuado para aproximar el verdadero valor. En este apartado se puede dar respuesta a la pregunta de si se puede medir con los datos de una sola muestra el error debido al muestreo para cada método. Destaca la cuestión conocida de la dificultad de estimar el error en el caso del muestreo sistemático.

1.6.3. VARIANZA DEL ESTIMADOR DE LA VARIANZA

Para un marco determinado es posible calcular para cada muestra el estimador de la varianza que cada método propone. Dado el espacio muestral completo podremos obtener su varianza. En la medida que el método de selección disponga a su vez de un método de estimación del error, con los datos de una sola muestra, que presente estabilidad podremos estar más confiados al analizar los resultados de una sola muestra. Esto es si, de obtener una muestra a obtener otra, el estimador del error va a diferir mucho entonces denominaremos al método inestable.

Tabla 3.- Varianza del estimador de la varianza.

MÉTODOS DE SELECCIÓN	Vvc[10]
Sistemático intervalo cte. (SIC)	443.17
Sistemático intervalo var. 1 (SIV1)	959.69
Sistemático intervalo var. 2 (SIV2)	573.93
Sistemático centrado (CLS)	52.32
Sistemático centrado int. var. 1 (CLSIV1)	133.78
Sistemático centrado int. var. 2 (CLSIV2)	185.21
Sistemático corrección de Yates (CY)	443.17
Sistemático equilibrado (BSS)	716.62
Sistemático modificado (MSS)	437.56
Prob. desiguales con reposición y ppt (CRPPT)	336.55
Prob. desiguales sin reposición y ppt (SRPPT)	306.15
Prob. desiguales con reposición parcial y ppt (CRPPPT)	286.72

[10] Vvc es la varianza del estimador de la varianza.

En este apartado se puede dar respuesta a la pregunta de si se puede tener confianza en la estabilidad de la medición, con los datos de una sola muestra, del error debido al muestreo para cada método.

1.7. MODELOS DE SUPERPOBLACIÓN

Una determinada población no deja de ser un hecho fortuito, una realización de un fenómeno que bajo ciertas circunstancias produce un resultado concreto. Basar un análisis de los errores de muestreo en una población circunstancial puede introducir irregularidades que desvirtúen el estudio. Para poder utilizar la información que proporciona un marco poblacional asegurándonos de introducir en el análisis una garantía de generalidad.

El concepto de superpoblación tiene una larga historia en la literatura sobre muestreo: Cochran(1939,1946), Deming y Stephan(1941), Madow y Madow(1944). Este enfoque fue introducido para comparar varianzas de métodos de selección alternativos. Se considera que la población finita sobre la que estamos trabajando es a su vez una muestra aleatoria de una población más general. Por tanto las demostraciones se llevan a cabo sobre una población patrón o superpoblación y no sobre una población finita concreta. Un modelo general que se utilizará al estudiar los métodos de probabilidades desiguales consiste en

$X_i = B M_i + e_i$ con $i = 1, 2, 3, ..., N$ donde e_i es una variable aleatoria. Por E^* representamos el valor esperado sobre todas las posibles poblaciones finitas que hipotéticamente pueden deducirse del modelo, condicionadas a un conjunto fijo de M_i valores.

$E^* (e_i / M_i) = 0$ $E^*_{ij} (e_i e_j / M_i M_j) = 0$ $E^*(e_i^2 (M_i)) = a M_i^g$

con a y g parámetros con valores empíricos que satisfacen las condiciones: $a > 0$, $1 \leq g \leq 2$

Un modelo se puede definir como una clase de distribución. Las especificaciones de esta clase pueden ir desde una formulación sobria, prescribiendo sólo ciertas características de la media, varianza y covarianzas, a una situación donde el nivel de detalle en la especificación sea muy alto.

Siguiendo a Cassel(1977) podemos ver que el punto sobresaliente del análisis estadístico bajo un enfoque de superpoblación es que la población finito, $X=(X_1, X_2 ... X_N)$ de la que se requiere información y que se define como población objetivo, es el resultado de un vector aleatorio finita, $X=(X_1, X_2 ... X_N)$ caracterizado por una distribución que se supone conocida. En muchas situaciones es natural que el modelo resuma y formalice el

conocimiento a priori disponible sobre la población. Este conocimiento puede estar basado en la experiencia o en una creencia subjetiva.

Se pueden distinguir dos enfoques en cuanto a los supuestos utilizados en la inferencia bajo modelos:

1.- Modelos que utilizan «instrumentos clásicos» (en el sentido no Bayesiano) de inferencia. Un supuesto típico es que el vector aleatorio tiene una distribución con parámetros desconocidos que deben ser previamente estimados.

2.- Modelos que utilizan «instrumentos Bayesianos». Así los parámetros desconocidos se estiman en base a distribuciones a priori.

El concepto de superpoblación puede tener distintas interpretaciones: así, los modelos de superpoblación se aplican a diferentes objetivos y pueden presentarse conforme la siguiente clasificación:

1.- La población finita es realmente una realización de un universo mayor. Esta es la idea de superpoblación en su forma más pura.

2.- El vector aleatorio y su distribución se modela o maqueta para describir un mecanismo o proceso aleatorio que se da en el mundo real. Desde este punto de vista se utilizan frecuentemente en econometría y sociometría.

3.- La distribución se considera a priori reflejando creencias subjetivas.

4.- La distribución, no se asocia ni a procesos del mundo real ni a creencias subjetivas, se utiliza como un instrumento matemático para obtener derivaciones teóricas.

5.- Puede ser útil como instrumento para el tratamiento de los errores ajenos al muestreo en el diseño de encuestas.

En este trabajo los modelos se utilizarán para comparar métodos de selección desde un punto de vista más general que el de una población finita concreta. Si estamos analizando una determinada población finita y se aplican diferentes métodos podemos clasificar estos métodos de menor a mayor varianza. Normalmente cuando los métodos presentan variaciones acusadas entre ellos, el repetir el cálculo de varianzas para otra población similar, con los mismos parámetros y estructura, no proporcionará resultados diferentes. Sin

embargo, si los métodos presentan variaciones menos acusadas y la población, al cambiarla por otra tiene un componente aleatorio importante, a pesar de tener características parecidas, entonces los resultados de los métodos pueden variar considerablemente. Se puede seguir esta parte con un ejemplo: supondremos que estamos diseñando la selección de una muestra con una población marco que presenta una tendencia lineal con heterocedastícidad. En los siguientes apartados vamos a mostrar algunos resultados obtenidos al calcular la varianza para una serie de métodos que estamos evaluando para diferentes realizaciones de esa población.

1.7.1 ALGUNOS RESULTADOS

Si realizamos, para la población de Cantabria analizada en el ejemplo anterior, un ajuste matemático, mediante interpolación y ajustamos el término de error del modelo para que represente la población lo más fielmente posible tendremos la expresión:

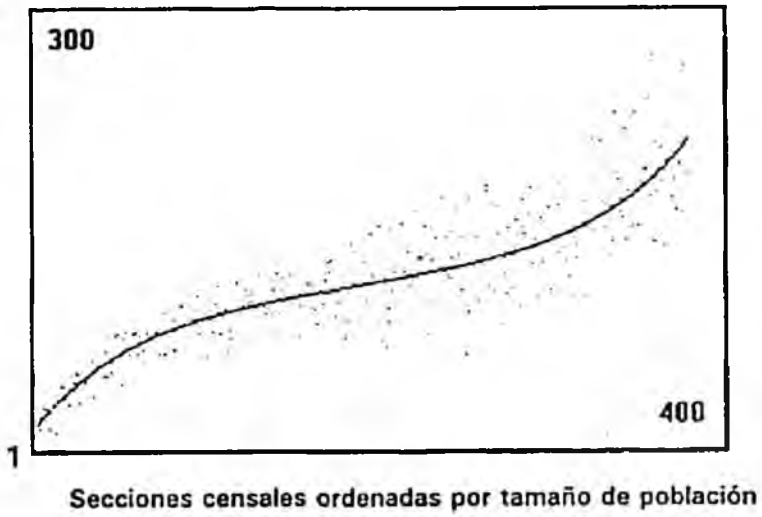
$$Xi = 1.8335E(+01) + 1.0158E(+00)i + -7.1831E(-03)*i^2 + 3.0846E(-05)i^3 + -$$

$$+ -7.0717E(-08)i^4 + 7.0852E(-11)i^5 + -e$$

donde e es el término de error que recoge la dispersión del fenómeno debida al azar y el efecto que se observa en la población de heterocedastícidad. Este término se distribuye con E(e)=0 y Var(e)=5(1+i*0.01)

La representación gráfica de una realización de este experimento se puede observar a continuación: los ejes y las definiciones se corresponden con una formalización matemática de la población de personas paradas en cada una de las 400 secciones censales ordenadas por tamaño de población.

Gráfico 2.- Variable de estudio para una superpoblación.[11]



Si generamos esta misma población un número elevado de veces y en cada una de ellas calculamos los parámetros que definen un diseño muestral tendremos los siguientes resultados:

[11] Fuente: elaboración propia.

Tabla 4.- Esperanza y varianza respecto del modelo de la varianza del estimador

MÉTODOS DE SELECCIÓN	E*Vp[12]	V*Vp[13]	CVS.	E* +(2*D*)
Sistemático intervalo cte. (SIC)	17.54	33.46	0.33	29.11
Sistemático intervalo var. 1 (SIV1)	11.84	15.58	0.33	19.74
Sistemático intervalo var. 2 (SIV2)	12.97	16.89	0.32	21.19
Sistemático centrado (CLS)	12.00	215.56	1.22	41.36
Sistemático centrado int. var. 1 (CLSIV1)	11.85	141.14	1.00	35.62
Sistemático centrado int. var. 2 (CLSIV2)	12.05	168.56	1.06	37.63
Sistemático corrección de Yates (CY)	12.97	20.56	0.35	22.04
Sistemático equilibrado (BSS)	12.25	19.11	0.36	20.99
Sistemático modificado (MSS)	11.68	17.67	0.36	20.09
Prob. desiguales con reposición y ppt (CRPPT)	12.84	13.86	0.29	20.28
Prob. desiguales sin reposición y ppt (SRPPT)	12.43	14.01	0.30	19.92
Prob. desiguales con reposición parcial y ppt (CRPPPT)	12.81	20.78	0.36	21.93

El principal resultado que observamos en esta tabla es el similar valor esperado de los métodos considerados y, en contraste la elevada variabilidad de los métodos centrados. Esto quiere decir que si bien el error esperado, al analizar varias poblaciones generadas de forma aleatoria es similar, éste varía mucho de unas poblaciones a otras. Por tanto, para esta población marco, la precisión respecto del modelo de los métodos centrados es pequeña en comparación con los otros métodos.

[12] E*Vp es la esperanza respecto del modelo de la varianza del estimador.

[13] V*Vp es la varianza respecto del modelo de la varianza del estimador.

Otro resultado que se puede obtener con el modelo es

Tabla 5.- Esperanza y varianza respecto del modelo del estimador de la varianza.

MÉTODOS DE SELECCIÓN	E*Vc[14]	V*Vc[15]	CVS.%	E* + 2D*
Sistemático intervalo cte. (SIC)	59.66	8.01	0.05	65.32
Sistemático intervalo var. 1 (SIV1)	59.81	8.22	0.05	65.55
Sistemático intervalo var. 2 (SIV2)	59.85	8.01	0.05	65.51
Sistemático centrado (CLS)	59.63	107.50	0.17	80.36
Sistemático centrado int. var. 1 (CLSIV1)	57.02	88.69	0.17	75.85
Sistemático centrado int. var. 2 (CLSIV2)	59.37	89.09	0.16	78.24
Sistemático corrección de Yates (CY)	59.66	8.01	0.05	65.32
Sistemático equilibrado (BSS)	59.88	8.20	0.05	65.60
Sistemático modificado (MSS)	59.88	8.21	0.05	65.61
Prob. desiguales con reposición y ppt (CRPPT)	13.43	4.62	0.16	17.73
Prob. desiguales sin reposición y ppt (SRPPT)	13.03	4.14	0.16	17.10
Prob. desiguales con reposición parcial y ppt (CRPPPT)	12.98	4.78	0.17	17.36

De esta tabla destacamos el valor poco acurado del estimador de la varianza que es posible esperar en los métodos sistemáticos. Se ha empleado para estimar la varianza en estos métodos sistemáticos la técnica del lazo, consistente en unir dos unidades contiguas para poder estimar simulando un muestreo estratificado con dos unidades por estrato. En contraste destaca la acuracidad de los métodos con probabilidades desiguales al estimar la varianza del estimador.

[14] E*Vc es la esperanza respecto del modelo del estimador de la varianza

[15] V*Vc es la varianza respecto del modelo del estimador de la varianza.

Y por último, en cuanto a estabilidad del estimador de la varianza tenemos el siguiente resultado:

Tabla 6.- Esperanza y varianza respecto del modelo de la varianza del estimador de la varianza.

MÉTODOS DE SELECCIÓN	E*VVc[16]	V*VVc[17]	CVS.	E* + 2D*
Sistemático intervalo cte. (SIC)	162.67	067.18	0.34	273.43
Sistemático intervalo var. 1 (SIV1)	165.05	3402.43	0.35	281.71
Sistemático intervalo var. 2 (SIV2)	165.79	3527.45	0.36	284.58
Sistemático centrado (CLS)	87.26	15700.22	1.44	337.87
Sistemático centrado int. var. 1 (CLSIV1)	72.57	16116.66	1.75	326.47
Sistemático centrado int. var. 2 (CLSIV2)	104.50	24664.57	1.50	418.60
Sistemático corrección de Yates (CY)	162.67	3067.18	0.34	273.43
Sistemático equilibrado (BSS)	165.05	3783.18	0.37	288.06
Sistemático modificado (MSS)	215.12	5171.71	0.33	358.95
Prob. desiguales con reposición y ppt (CRPPT)	48.11	487.22	0.46	92.26
Prob. desiguales sin reposición y ppt (SRPPT)	42.23	502.11	0.53	87.05
Prob. desiguales con reposición parcial y ppt (CRPPPT)	48.05	820.35	0.60	105.34

Observamos que, en cuanto a valor esperado de la estabilidad del estimador de la varianza, los métodos centrados vuelven a presentar una elevada inestabilidad. Destaca también la inestabilidad del método modificado.

[16] E*VVc es la esperanza respecto del modelo de la varianza del estimador de la varianza

[17] V*VVc es la varianza respecto del modelo de la varianza del estimador de la varianza.

1.8. FUNCIÓN DE CORRELACIÓN INTRACLÁSICA: CORRELOGRAMA

En el muestreo sistemático no se produce la relación conocida en el muestreo aleatorio entre la varianza del estimador y el tamaño de muestra. Esto es debido a que la varianza depende de la varianza poblacional y del valor que tome el coeficiente de correlación intraclásica. El cual a su vez depende de la relación que exista entre los grupos y por tanto depende del número de clases o grupos que se formen.

Con la aplicación POSDEM podemos obtener, para una determinada población, los valores del coeficiente de correlación para distintos valores de tamaño de muestra. Por correlograma se entiende la representación gráfica del coeficiente de correlación intraclásica para diferentes tamaños de muestra. Por ejemplo, para la población que hemos considerado antes tendríamos la siguiente tabla y gráfico:

Tabla 7.- Coeficiente de correlación intraclásica y otros.

n[18]	Vp(m)[19]	CV_m%	[20]Vp(st)	CV_st%	Vp(sr)[21]	CV_sr%	Corr[22]	-1/n	-1/(n - 1)	D[23]	N[24]	S(x)[25]
2	275.95	18%	454.76	23%	748.06	30%	-0.632	-0.50	-1.00	Si	400	200
4	113.66	12%	171.26	14%	372.15	21%	-0.232	-0.25	-0.33	No	400	100
8	57.20	8%	71.52	9%	184.2	15%	-0.099	-0.13	-0.14	No	400	50
10	56.75	8%	55.45	8%	146.61	13%	-0.069	-0.10	-0.11	No	400	40
16	22.88	5%	33.76	6%	90.22	10%	-0.050	-0.06	-0.07	No	400	25
20	21.97	5%	26.75	6%	71.42	9%	-0.037	-0.05	-0.05	No	400	20
25	12.05	4%	21.08	5%	56.39	8%	-0.035	-0.04	-0.04	No	400	16
40	5.89	3%	12.91	4%	33.83	6%	-0.022	-0.03	-0.03	No	400	10
50	8.35	3%	9.86	3%	26.31	6%	-0.015	-0.02	-0.02	No	400	8
80	1.36	1%	5.71	3%	15.04	4%	-0.012	-0.01	-0.01	No	400	5
100	2.92	2%	4.35	2%	11.28	4%	-0.008	-0.01	-0.01	No	400	4
200	0.17	0%	1.50	1%	3.76	2%	-0.005	-0.01	-0.01	No	400	2

El método empleado, como ejemplo, ha sido el muestreo modificado. En la tabla tenemos el valor del coeficiente de correlación intraciásica para diferentes combinaciones de tamaños de muestra, al variar éste varía a su vez el número de reiteraciones o de grupos que se deben formar en la población para obtener una muestra sistemática. De esta manera podemos observar la forma del correlograma y comprobar que para el tamaño de muestra

[18] n es el tamaño de muestra

[19] Vp(m) es la varianza del estimador utilizando muestreo sistemático modificado

[20] Vp(st) es la varianza del estimador utilizando muestreo estratificado con una unidad por estrato

[21] Vp(sr) es la varianza del estimador utilizando muestreo aleatorio sin reposición

[22] Corr es el coeficiente de correlación intraclásica

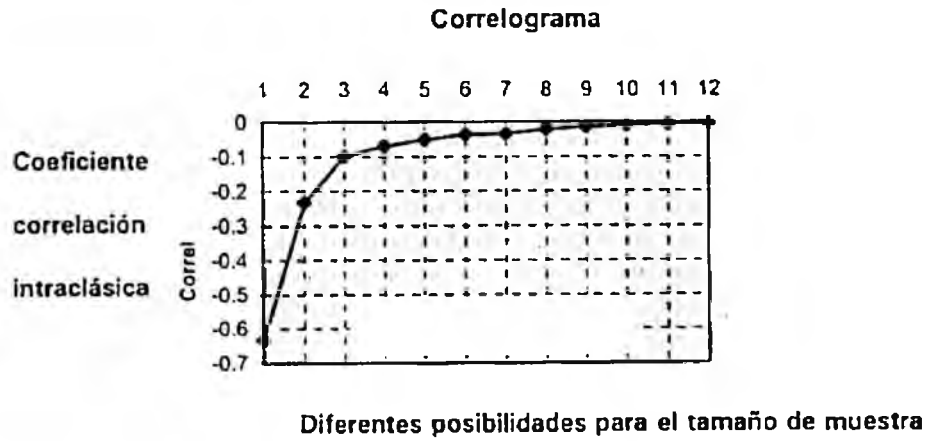
[23] D es un parámetro que significa si corr se encuentra entre los límites -1/n y -1/(n-1)

[24] N es el tamaño de la población en número de unidades

[25] S(x) es el número de muestras que forman el espacio muestral dado el método de selección

considerado anteriormente se verifica que la varianza esta situada en un valor próximo a la cota superior marcada por el modelo de superpoblación. No es el mismo puesto que el correlograma lo estamos obteniendo de una realización concreta de esa estructura poblacional. Se puede observar también como en el muestreo sistemático la relación entre tamaño de muestra y varianza poblacional no es la misma que en el muestreo aleatorio, en este caso aumentos de tamaño de muestra pueden suponer aumentos en la varianza del estimador.

Gráfico 3.- Correlación intraclásica y distintos tamaños de muestra.[26]



[26] En el eje de abcisas se representan las doce situaciones consideradas en la tabla en cuanto a tamaño de muestra

Por otra parte, en este trabajo se ha puesto el coeficiente de correlación intraclásica en relación con las expresiones -1/(n-1) y -1/n. Esto es debido a que, cuando el coeficiente de correlación intraclásica se encuentra entre esos límites, se puede demostrar fácilmente que la varianza del estimador se puede calcular como el resultado del cociente entre la varianza del muestreo sin reposición y el tamaño de muestra.

1.9. RESEÑA HISTÓRICA DE LA TEORÍA DEL MUESTREO SISTEMÁTICO

El muestreo sistemático ha sido considerado con detalle por W.G. Madow y L.H. Madow (1944), L.H. Madow(1946), Cochran (1946), W.G. Madow(1949,1953) y Lahiri (1954). Reseñas de los trabajos realizados en este campo se han llevado a cabo por Yates (1948), Bucklan (1951) e Iachan (1982). Aplicaciones del muestreo sistemático a las encuestas forestales se han ilustrado por Hasel (1938), Finney (1948) y Nair y Bhargava (1951). La utilización de muestreo sistemático en las encuestas sobre pesca ha sido mostrado por Sukhatme, Panse y Sastry (1958)

Los primeros trabajos que pusieron de manifiesto la importancia del muestreo sistemático surgieron en el campo de los estudios forestales y fueron debidos a Hasel y Osborne en 1938 y 1942. Posteriormente los Madow publicaron sus trabajos sobre la teoría del muestreo sistemático en 1944. Este artículo, junto con la extensión a modelos de superpoblación de Cochran en 1946,

constituyen la base sobre la que se han ido agregando las contribuciones posteriores. Entre éstas hay que citar los trabajos de Quenouille y Yates en 1949.

En 1952, Buckland lleva a cabo la primera revisión de los trabajos que hasta esa fecha habían ido configurando el campo del muestreo sistemático.

En 1965, Sheti, presenta el método equilibrado. En 1968 Singh, Jindal y Garg publican el método modificado. Puede encontrarse una descripción completa con aplicación a poblaciones naturales en el libro de Murthy de 1968. En 1975, Bellhouse y Rao, desarrollan los resultados de estos métodos en el marco de los modelos de superpoblación para diferentes esquemas de población.

En 1982, Iachan lleva a cabo la segunda revisión de los trabajos que desde 1952 hasta la fecha se habían incorporado a la teoría y práctica del muestreo sistemático. En el caso de aplicación a poblaciones naturales es interesante el libro de Krishnaiah y Rao de 1988, donde pueden estudiarse dos capítulos sobre muestreo sistemático elaborados, respectivamente, por Bellhouse y Murthy y Rao.

1.10. MÉTODOS INTENSIVOS POR ORDENADOR: BOOTSTRAP Y JACKKNIFE

Siguiendo a Efron, vemos que la libertad que producen las posibilidades de cómputo, entre otras, es la liberalización del factor limitante que obliga a concentrar los esfuerzos sobre medidas estadísticas de propiedades teóricas susceptibles de ser analizadas matemáticamente. Por ejemplo, el problema de la confianza, precisión de una estimación como el coeficiente de correlación, que fue resuelto por Fisher en 1915 con los procedimientos teóricos establecidos en aquel momento, es, ahora, con los métodos intensivos de ordenador y concretamente con el procedimiento debido a Efron denominado de muestras autogeneradas o "Bootstrap", cuando es posible determinar este nivel de confianza para el coeficiente de correlación de manera más exacta. No obstante, existen todavía medidas ampliamente utilizadas como es el caso de las "componentes principales" utilizadas para resumir un conjunto más amplio de información. Establecer la confianza de estas

estimaciones en los términos clásicos de la teoría del muestreo en poblaciones finitas presenta un nivel de dificultad prácticamente insalvable a no ser por los procedimientos intensivos de ordenador.

Existen otros métodos estadísticos que se fundan en la potencia de los ordenadores: se ha realizado una referencia al método denominado "Bootstrap", pero existen otros como el método "Jackknife" debido a Quenouille (1949) y a Tuckey (1950), el método de validación reciproca o el conocido por replicación repetitiva equilibrada. Todos estos métodos difieren entre sí y presentan distintas posibilidades, pero tienen como componente común que generan datos ficticios a partir de datos originales y calculan luego la variabilidad real de un estadístico a partir de su variabilidad sobre el conjunto de datos ficticios. Las diferencias entre ellos residen en los procedimientos seguidos para engendrar tales datos ficticios.

Podemos utilizar un ejemplo de los procedimientos "Bootstrap" y "Jackknife" resueltos con la aplicación POSDEM. El ejemplo para "Bootstrap" lo tomamos de Efron. Se considera que se está investigando en 15 Universidades la relación, a través del coeficiente de correlación, de dos indicadores del nivel de rendimiento académico. Este coeficiente toma, para esta muestra concreta de 15

universidades el valor 0.776, el problema reside en estimar el grado de confianza que merece la estimación de este coeficiente. Esto es, en que medida, si obtenemos otra muestra con otras 15 universidades, podemos esperar que el comportamiento del coeficiente de correlación se encuentre entre ciertos límites.

El método "Bootstrap" utiliza las quince unidades muestrales para generar una población ficticia con el siguiente mecanismo: selecciona la primera unidad y la repite un número elevado de veces, (Efron cita un millón de veces). Nosotros consideraremos su repetición cien veces. EL mismo procedimiento se aplica para la segunda unidad, y así sucesivamente. De esta forma tendremos una población ficticia de 1500 unidades. De ahí se puede generar una representación del espacio muestral que se denomina muestras autogeneradas o muestras "Bootstrap"; el estudio de esta distribución permite conocer la precisión del estadístico que estamos considerando.

EL método "Jackknife", navaja de Jack, fue acuñado por Tukey, con la intención de sugerir que el método es un instrumento estadístico con utilidad general, como la navaja. Por ejemplo, dadas las 15 universidades, el método consiste en ir suprimiendo cada vez

una observación del conjunto inicial de datos, y recalcular el estadístico que nos interesa para cada uno de los conjuntos de datos así truncados.

1.11. CONCLUSIONES

En la práctica de las encuestas por muestreo uno de los problemas fundamentales con que se enfrenta el diseñador es el de disponer de un sistema de selección de unidades muestrales a partir de una población marco u objetivo que haga óptimas las condiciones del diseño. Es muy habitual el realizar un diseño bietápico con selección de unidades de primera etapa con probabilidades proporcionales al tamaño y, posteriormente o en segunda etapa, una selección sistemática de unidades últimas con probabilidades iguales. Es también frecuente que el número de unidades últimas sea constante, de forma que las probabilidades de selección permanecen constantes para cada unidad de la población.

Una vez obtenidas la muestra y tabulados sus principales resultados se procede a la formación de una colección de tablas con sus errores de muestreo. En encuestas más elaboradas incluso se publican resultados sobre la evaluación de los errores ajenos al muestreo. Puede verse la publicación de la Encuesta de Población Activa, elaborada por el INE.

Los métodos de selección se utilizan en base a consideraciones teóricas que demuestran las ventajas de unos métodos respecto de otros. No obstante siempre se requiere el disponer o el realizar algún tipo de supuesto sobre la estructura de la población. Por ejemplo en el muestreo sistemático se supone que no existen periodicidades que puedan afectar al proceso de estimación. En el muestreo con probabilidades proporcionales se supone que las variables objeto de estudio están relacionadas con la variable utilizada para calcular las probabilidades de selección. Estos supuestos suelen ser lo suficientemente razonables como para darlos por buenos.

Sin embargo, cuando se dispone de abundante información auxiliar sobre la población es conveniente el estudiar la estructura de esa población en relación con los métodos de selección propuestos para mejorar la precisión de los estimadores y para seleccionar el mejor método posible. Dentro de estos estudios entra el considerar la población como una muestra de una población más general, superpoblación, de la cual la población natural que se estudia no es más que una de las posibles que se ha formado de acuerdo con unas fuerzas que podrían haber formado igualmente otra población similar pero aleatoriamente diferente. Este tipo de estudios tiene gran

tradición en la teoría del muestreo en poblaciones finitas desde que Cochran lo aplicase para estudiar la comparación de varianzas debidas a métodos diferentes. Sin embargo estos estudios no se suelen llevar a cabo para preparar el diseño de encuestas concretas. En nuestro caso se ha procedido a estudiar la distribución de la población en base a ciertas características censales. Posteriormente se ha modelado esta población natural a población aleatoria y se han comparado las varianzas esperadas y la varianza de las varianzas, ambas respecto del modelo. Para terminar se sugiere un diseño de selección y de estimación que supone una mejora sustancial respecto de sus alternativos para esta encuesta concreta.

Por tanto, entre las aportaciones de esta investigación se encuentran:

En primer lugar, la materialización de un método nuevo para determinar diseños de selección ajustados a estructuras poblacionales determinadas, bajo el enfoque de modelos de superpoblación. Esta materialización se concreta en el programa de ordenador POSDEM, programa para optimizar la selección en el diseño de encuestas por muestreo.

En segundo lugar se propone un nuevo método sistemático de selección con intervalo de muestreo variable. Este método elimina la tendencia resultado de la ordenación de las unidades poblacionales, presentando el caso de varianza cero para poblaciones hipotéticas $X_i = i$ con $i = 1,2,3...N$ cuando se verifica la condición de que el tamaño de muestra es igual a la raíz cuadrada de la población. En este caso presenta importantes mejoras en precisión respecto del muestreo aleatorio simple y el muestreo sistemático con intervalo constante. En relación con los otros métodos considerados, equilibrado, modificado, centrado y corregido presenta un comportamiento similar, con la condición de la raíz, que depende de la estructura concreta de la población. El caso donde el método supone una mejora importante respecto a estos últimos métodos citados es en poblaciones que presentan una periodicidad cíclica. Poblaciones que se dan en la práctica muestral y para las que los libros de texto consultados coinciden en tomar grandes precauciones al aplicar los métodos tradicionales de muestreo sistemático. Con los métodos tradicionales la única forma de evitar los perjuicios que supone la existencia de ciclos en la población era conocerlos perfectamente y poder contrarrestarlos en el diseño. Con el método

de intervalo variable incluso en el caso de no tener información sobre el comportamiento cíclico éste tiende a quedar eliminado por la misma definición del procedimiento. Este resultado nos ha llevado a definir otro método de muestreo sistemático con intervalo variable, que hemos denominado intervalo variable II, donde hemos planteado el razonamiento siguiente: si en la teoría del muestreo sistemático clásico la ordenación se aconsejaba para que el diagrama de puntos tenga una tendencia, cuando se aplica el muestreo sistemático con intervalo variable es aconsejable ordenar la población de forma que los valores de las unidades sigan una curva cíclica. La utilidad, por mejora en precisión de este segundo método depende también de la estructura concreta de la población que se utilice como marco. El desarrollo de este método podría ser una futura línea de investigación. Otro método que permite introducir la novedad del intervalo variable es el método centrado con dicho intervalo en sus dos versiones de intervalo variable I y II. En esta línea hemos encontrado resultados interesantes en cuanto al comportamiento de los nuevos métodos propuestos, en especial del muestreo sistemático centrado con intervalo variable, frente a cambios en las especificaciones del

modelo. Esto nos ha llevado a recomendar, para ciertas clases de poblaciones, este método sobre los restantes.

Y, por último, se incorporan a los estudios empíricos el esquema teórico de modelos de superpoblación. Es posible trabajar empíricamente con diferentes tipos de modelos según sea el ajuste a los datos que se estén considerando. E incluso definir modelos por tramos de población y tratarlos de manera conjunta. Como un subproducto de la investigación se proporciona un programa de ordenador, POSDEM, con bastantes funcionalidades, entre las que se encuentran además de las citadas, las necesarias para realizar experimentaciones autogeneradas, "Bootstrap" y "Jackknife"

Entre las líneas de investigación pendientes que observamos en este trabajo se encuentra el desarrollo de un módulo de aprendizaje interno para que el ordenador pudiera ir incorporando información dentro de un esquema de inteligencia artificial. Desarrollo este último que queda fuera del área propiamente de lo que es la teoría y práctica del muestreo en poblaciones finitas. Otras ampliaciones podrían venir dadas por la consideración de otros parámetros, como la media armónica, geométrica, medidas de asociación, de apuntamiento, etc. . Incorporar este tipo de parámetros, desde el punto de vista de la

programación de ordenador no supone ninguna dificultad. Únicamente no se ha llevado a cabo por una cuestión de simplicidad. No obstante, este tipo de estudio sobre otros parámetros diferentes a la media o al total, puede dar lugar a estudios empíricos que habitualmente quedan fuera de su consideración teórica por su dificultad de desarrollo. Y que, sin embargo, al utilizar las facilidades de cómputo de los ordenadores y de la aplicación POSDEM podrían llevarse a cabo.

CAPITULO II

INTRODUCCIÓN A POSDEM

2.1. ¿ QUÉ PERMITE LA APLICACIÓN POSDEM?

La aplicación POSDEM se ha diseñado de forma que permite, en el diseño de encuestas por muestreo probabilístico, «optimizar la selección» de las unidades que van a formar parte de una muestra. Utiliza diferentes métodos de selección y poblaciones definidas por el usuario. Básicamente esta aplicación informática tiene tres propósitos o puede ser utilizada desde tres ópticas diferentes.

Por una parte, se puede utilizar como un «instrumento pedagógico»: bajo este punto de vista permite resolver problemas de selección de unidades muestrales en el diseño de encuestas con cualquier tipo de datos. En la misma línea pedagógica también permite trabajar con variables de una población objeto de estudio de la que se dispone de información auxiliar de tipo censal y que constituye el marco que se va utilizar. De esta forma es posible obtener el espacio muestral completo, ó bien una representación del espacio muestral; también permite obtener el espacio paramétrico y del error de muestreo, y resulta más sencillo comprender mejor ciertos aspectos de la teoría del muestreo en poblaciones finitas. Así, es posible estudiar problemas donde,

definido un cierto marco de unidades, se obtienen un conjunto de muestras probabilísticas con sus características, estimaciones y errores asociados; calculando, además, el estimador para cada muestra y los errores debidos al muestreo para diferentes métodos alternativos.

En segundo lugar constituye un «instrumento de investigación» empírica. Permite determinar qué método de selección es preferible para una determinada estructura poblacional, y cuáles son las relaciones entre la estructura poblacional y los procedimientos de muestreo utilizados. La aplicación facilita el trabajar con modelos de probabilidades iguales, desiguales; modelos de superpoblación y diseños complejos bietápicos.

En tercer lugar, se trata de una «herramienta del trabajo de campo» de una encuesta por muestreo. Dado un marco de unidades, y una vez elegido el método de muestreo que mejor se ajusta a esa estructura poblacional, entonces es posible determinar qué unidades de la población pertenecerán finalmente a la muestra que será investigada, y cuales serán visitadas en caso de ausencias, negativas u otras incidencias.

En resumen y desde esa triple óptica este programa será de «utilidad»: a alumnos de un curso de teoría y práctica de muestreo en

poblaciones finitas en sus dos posibles versiones de básico o avanzado; a profesores que deseen disponer de un instrumento para la investigación empírica del área de las encuestas por muestreo; y, a empresas, oficinas centrales de estadística, o profesionales de investigaciones por muestreo que deseen diseñar encuestas con una selección óptima de las unidades muestrales. Este instrumento, la aplicación POSDEM, permite realizar diseños de una manera personalizada a cada investigación e incorpora el conocimiento que el experto en muestreo tiene, no siendo necesario que la persona que efectivamente realiza el diseño tenga estos conocimientos.

Vamos a resumir lo que permite hacer POSDEM:

1.- El programa permite trabajar con «bases de datos» procedentes de ficheros externos en formatos Dbase, Paradox, Foxpro, y Acces. También permite generar variables definidas a medida y generar poblaciones aleatorias bajo una amplia gama de posibilidades: aleatoriedad, dispersión, heterocedastícidad, tendencia y ciclo. De forma que es posible simular una gran variedad de poblaciones que se pueden encontrar en la práctica de las encuestas por muestreo. Esta posibilidad se utiliza sobre todo dentro del enfoque de modelos de superpoblación.

2.- El programa tiene incorporados diecisiete «métodos de selección» y veinte métodos de estimación distintos. Una vez elegido el método y obtenidas las muestras, calcula para cada muestra, el estimador del total, de la media, o de la proporción, según se trate de datos cualitativos o cuantitativos; la varianza del estimador, su desviación típica, los límites de confianza al 95% y el coeficiente de variación. Calcula también, para ciertas opciones, los momentos de segundo, tercer y cuarto orden. Todos estos cálculos se realizan para cada muestra obtenida, y se pueden listar bajo diferentes formatos, con salidas a pantalla, a impresora, a disco o al portapapeles de Windows. Para una población tipo de 800 unidades calcula como resultado final y en una sola realización, sin tener en cuenta procesos intermedios o cálculos definidos específicamente, un total de 11.298 estimaciones diferentes. Dispone además de cien opciones específicas, todas ellas accesibles desde menú. Las más utilizadas dispuestas en teclas de función y teclas abreviadas (combinación de ctrl+letra). Presenta también una ayuda en línea de cada pantalla con su correspondiente glosario. Esta ayuda se presenta en formato hipertexto.

3.- De los procedimientos de selección, que utiliza la aplicación, tenemos un primer grupo con doce procedimientos que tienen

en común que la selección se realiza con «probabilidades iguales», y un segundo grupo de tres métodos con «probabilidades desiguales», proporcionales al valor de una variable auxiliar, generalmente el tamaño, que suele estar correlacionado con la variable de estudio.

4.- Las muestras obtenidas se utilizan a su vez para calcular el «error de muestreo». Esto es, el programa calcula la varianza, desviación típica, coeficiente de variación y límites de confianza para todos los estimadores calculados en cada muestra. Así, tenemos la posibilidad de estudiar la varianza de la media y también la varianza de la varianza o la varianza del momento de cuarto orden, entre otros parámetros. Hay que destacar también que con este procedimiento, en la línea de los procedimientos "bootstrap" es posible calcular la varianza para estadísticos más sofisticados que la media o el total, como por ejemplo una componente principal.

5.- La aplicación utiliza por defecto «variables cuantitativas», aunque también permite procesar datos cualitativos. Y en cuanto a «estimadores», es posible elegir el tipo de estimador que vamos a utilizar: la media, el total, la proporción, el total de clase o incluso, si se dispone de la información adecuada, es posible utilizar estimadores mejorados de regresión y de razón.

6.- Dentro de la filosofía del programa un punto básico es poder realizar «comparaciones entre diferentes métodos de selección», para poder elegir el más conveniente a cierta estructura poblacional. El programa permite estudiar los resultados de cada experimento en forma de listados o gráficamente. En cuanto a la estructura de población, esta puede estar definida por una población ficticia, una población natural, observada en la práctica, o una "superpoblación", producto de un modelo.

7.- Los «modelos de superpoblación» es uno de los puntos fuertes del programa puesto que permite definir modelos muy complejos, por tramos, con distintas formas y características en cuanto a término de error, heterocedasticidad, concavidad, convexidad o componente cíclico. Admite ajustes con polinomios ortogonales.

8.- Permite también realizar los cálculos necesarios para representar gráficamente el coeficiente de correlación intraclásica y la varianza, mediante un «correlograma», definiendo los límites inferior y superior del tamaño de muestra. A su vez permite realizar una «descomposición de la varianza», distinguiendo por fuente de variación entre muestras o dentro de muestras.

9.- Para análisis de tipo multivariante se ha incorporado un «módulo de componentes principales» que permite obtener la componente de un determinado conjunto de variables para su utilización en el diseño de la encuesta bien como variable de estudio, bien como variable auxiliar o de ordenación , en función del diseño que se este realizando.

10.- Permite, por último realizar «diseños polietápicos». Así, en primer lugar es posible obtener las unidades que formarán la muestra de unidades primarias sobre las que a su vez se realizará un nuevo muestreo, hasta conseguir determinar las unidades últimas de estudio. La aplicación tiene implementado un procedimiento para obtener el error de muestreo en diseños bietápicos.

2.2 ¿ CÓMO SE ENCUADRA DENTRO DEL MARCO ESTADÍSTICO E INFORMÁTICO ?

Dentro del marco estrictamente informático, la simulación de situaciones estocásticas por ordenador para determinar soluciones óptimas de ciertos problemas tiene una utilidad innegable. Desde el comienzo del cálculo computarizado se ha prestado mucha atención al proceso de entrada de datos y posterior tratamiento de cálculo. Sin embargo, tradicionalmente ha existido una laguna precisamente en la fase anterior a la entrada de datos. Esto es, en el momento de determinar que unidades se van a estudiar, en definitiva, en procedimientos que simulasen el proceso de muestreo. Se trata de utilizar el ordenador para un etapa previa, en lo que forma el diseño de una encuesta, a la entrada de datos, con el fin de determinar que unidades deben formar parte de una muestra probabilística. Por tanto, esta aplicación es previa a aquellas de entrada de datos, de tabulación y cálculo, o de análisis estadístico.

Dentro del marco estadístico, con esta aplicación se ha pretendido realizar un programa que simule situaciones que permitan

elegir el óptimo entre varios diseños de selección muestrales. Se podría incluso ir configurando un diseño especifico óptimo para cada estrato. En el muestreo de poblaciones finitas las técnicas pueden catalogarse en orden de dificultad desde el muestreo con reposición al muestreo polietápico con diferentes esquemas de muestreo. Se puede, a su vez, distinguir dos grandes grupos en función de si las probabilidades de selección de cada unidad en la población es igual o desigual a las restantes. En esta aplicación se han implementado procedimientos de selección de muestras desde el muestreo con reposición al muestreo de conglomerados, pasando por el muestreo sistemático y el muestreo estratificado. Igualmente se han implementado métodos para el caso de probabilidades desiguales con, sin y con reposición parcial. Existen además variantes para el cálculo de los errores debidos al muestreo como es el caso de que el orden de la población sea aleatorio, la técnica del lazo, las pseudoreiteraciones con semimuestras, una variante para el muestreo de conglomerados utilizando la varianza intraconglomerados o métodos autogenerados, "jackknife", "bootstrap".

2.3. PROGRAMAS DE ORDENADOR PARA EL ANÁLISIS DE DATOS PROCEDENTES DE ENCUESTAS COMPLEJAS

Dentro de los programas de ordenador aplicados a la teoría del muestreo en poblaciones finitas puede encontrarse una buena introducción en el artículo de Jim Lepkowski y Judy Bowles, "Sampling error software for personal computers". En este artículo se realiza un resumen de las razones por las que es necesario un software especial para analizar datos procedentes de encuestas y se describen ocho paquetes informáticos para PC. Los programas analizados en este artículo son Cenvar, Clusters, Epiinf, Pccarp, Stata, Sudaan y Wervarpc. Estos programas tienen en común que permiten analizar convenientemente, al tener en cuenta las especificaciones del diseño, los datos procedentes de encuestas en diseños complejos.

2.4. PRINCIPALES PANTALLAS DE LA APLICACIÓN

Las principales pantallas de la aplicación se muestran en el anexo.

En la primera pantalla de este anexo podemos ver en su primera línea el titulo y descripción del programa. En la segunda aparecen una serie de palabras que representa opciones de un menú, concretamente: datos, muestreo, cálculos, listados, gráficos, opciones, ventana y ayuda. Cada una de estas opciones puede estar activada o desactivada dependiendo de si en ese momento tiene o no sentido su ejecución. En el caso de estar inactiva aparecerá escrita en gris claro y no podrá ser activada por el usuario. En el caso de estar activa aparecerá en una letra negra y el usuario podrá, al seleccionarla optar entre varias posibilidades de ejecución. El resto de la pantalla es un conjunto de ventanas unas contienen listados o resultados de cálculos y otras gráficos. Esta pantalla muestra lo que podríamos denominar una «pantalla obtenida por un utilizador experto de POSDEM». En ella hemos seleccionado con datos obtenidos de diferentes fuentes, un listado de ficha técnica, que muestra las

principales definiciones que suelen aparecer en las fichas de encuestas por muestreo: método empleado, error de muestreo, población objetivo, parámetros poblacionales para la variable de estudio e información auxiliar utilizada. Debajo de esta ventana aparece un gráfico con los valores de la variable analizada para toda la población marco. Otro gráfico, que se ha seleccionado para esta pantalla es el de las estimaciones y su error. En él podemos observar las estimaciones obtenidas, para un determinado método, como se sitúan alrededor de su valor esperado. También se observa en él, el error de muestreo asociado con cada estimación, con cada muestra, puesto que la círculo que señala el punto se calcula en base a la desviación típica del estimador. El listado inmediatamente siguiente muestra los valores obtenidos para diferentes tamaños de muestra, también para un método determinado, del error de muestreo correspondientes y de los valores que toma el coeficientes de correlación intraclásica. El gráfico que muestra como varían conjuntamente los coeficientes de correlación intraclásicos y el tamaño de muestra se denomina correlograma, y se ha incorporado también a esta primera pantalla. Por último, y sirviendo de base, aparece una ventana alargada, que muestra los resultados obtenidos

cuando se obtiene un modelo de una población natural con un esquema de superpoblación y se comparan las varianzas de diferentes métodos de selección. Este gráfico por similitud con el correlograma se ha denominado supergrama.

La segunda pantalla que puede observarse en el anexo es la misma pantalla anterior pero "limpia", esto es, tal y como aparecerá «la pantalla cuando se inicie una sesión». Ahora es más sencillo observar la línea de menú con sus opciones. Dentro de esta pantalla destacan tres elementos: una primera ventada de listados donde irán acumulándose los cálculos o listados que se vayan realizando durante el diseño. En el margen aparecen tres botones y una casilla en blanco. Los botones permiten mandar la salida que en ese momento aparece en la ventana de listados, bien a un fichero en disco, al portapapeles del sistema o a la impresora. Desde el portapapeles será posible transferir esta salida a cualquier procesado de textos. La ventana de gráficos tiene la misma funcionalidad que la de listados pero referida a gráficos. Y, por último, minimizado, tenemos en la línea inferior el editor de datos. Editor que veremos con más detalle en su pantalla correspondiente.

La siguiente pantalla muestra lo que obtendríamos al seleccionar la opción de "Datos". Así tendríamos cuatro posibilidades. En este punto conviene resaltar que es una práctica habitual de los programas en ventana el mantener ciertas formas normalizadas, de esta manera cuando detrás de una opción aparecen tres puntos suspensivos el usuario sabe que al seleccionarla el programa no ejecutará, todavía, ninguna orden directa sino que mostrará un cuadro de dialogo que permitirá al usuario especificar más su petición. Cuando en lugar de los tres punto aparece un triángulo, esto se interpreta como que al seleccionar esa opción tendremos la posibilidad de seleccionar otras más detalladas. Volviendo a la opción de datos, en primer lugar podremos seleccionar distintas alternativas: importar bases de datos, generar estructuras de población, editar los datos directamente por pantalla o bien realizar un análisis descriptivo de frecuencias, medidas o correlaciones. Este es un análisis exploratorio que puede realizarse con cualquier otro programa de análisis estadístico, pero que se ha incorporado para simplificar el procedimiento al usuario.

Si seleccionamos importar bases de datos, aparecerá un cuadro de dialogo, que permitirá indicar al programa donde se

encuentra el fichero que tiene los datos que deseamos procesar. Normalmente este fichero contendrá las unidades de la población con sus correspondientes variables. El formato de este fichero puede ser dbase, foxpro, paradox ó acces.

Una vez determinado un conjunto de datos, POSDEM muestra una pantalla con el "Editor de datos" donde podemos observar y editar, en su caso, los valores del fichero seleccionado anteriormente. En esta pantalla el programa necesita que el usuario determine, al menos, una variable de estudio. También puede fijar una variable de identificación, si los métodos que va a emplear lo precisan, deberá determinar una variable auxiliar, y , por último, si es necesario determinar una variable para ordenar la población. El editor de datos presenta también la posibilidad de suprimir o añadir registros, definir nuevas variables y filtrar datos. Estas dos últimas opciones dan lugar a dos nuevos cuadros de dialogo, que se muestran en el anexo , y que permiten realizar esas operaciones con una gran versatilidad. Es posible definir variables con estructuras muy complejas o seleccionar subconjuntos de datos que verifiquen varias condiciones.

Cuando, al contrario de trabajar con una población natural, se opta por generar una estructura poblacional, entonces POSDEM muestra el cuadro de dialogo que hemos denominado "Simulador de estructuras poblacionales". En éste cuadro el usuario puede determinar el tamaño de la población y su forma. Para esto podrá optar entre poblaciones con tendencia lineal, exponencial, potencial, y cíclicas, así como determinar los parámetros de término independiente, pendiente, media y varianza del componente aleatorio, grado de dispersión y grado de heterocedasticidad. Con este cuadro de dialogo es posible definir una estructura poblacional teórica, que se conoce en la práctica de una determinada investigación, o simplemente una población artificial para la que se desea analizar varios métodos.

La última pantalla que analizaremos en este capítulo es la que muestra el resultado de seleccionar la opción inicial de "Muestreo". En este caso el usuario podrá optar entre 6 alternativas disponibles. Las cuatro primeras hacen referencia a métodos de selección con probabilidades desiguales, la quinta a los métodos con probabilidades desiguales y la sexta al muestreo bietápico. Dentro de la alternativa de muestreo aleatorio simple se han considerado las

opciones con y sin reposición, para el muestreo sistemático se han considerado cinco tipos diferentes. Estos mismos métodos se han tenido en cuenta para el caso de muestras con dos conglomerados. Para el caso de muestreo con probabilidades desiguales se han considerado los esquemas con, sin y con reposición parcial. Y, por último dentro de la opción "Muestreo" se ha implementado la posibilidad de calcular el error de muestreo para ciertos casos de diseños en dos etapas.

CAPITULO III

UTILIDADES DE POSDEM

3.1. CÁLCULOS QUE ES POSIBLE REALIZAR CON POSDEM

En el menú de la aplicación, al seleccionar la opción de "Cálculos" aparecen diferentes alternativas, en primer lugar es posible obtener cálculos para cada muestra y cálculos para todas las muestras. Con la opción de "Cálculos para cada muestra" se obtiene para todas las muestras un listado donde figura el estimador media o proporción según los datos sean cuantitativos o cualitativos; la varianza del estimador; su desviación típica; los límites inferior y superior de confianza para el 95% de confianza; y el coeficiente de variación o error de muestreo en términos relativos. Con la opción de "Cálculos para todas las muestras" se obtiene un estudio para el conjunto de las reiteraciones consideradas que permite disponer de la media, varianza, desviación, límites y coeficiente de variación para cada uno de los parámetros estimados en cada muestra. Es posible estudiar la variabilidad que se presenta al repetir el proceso de muestreo de cualquiera de los parámetros que se investigan. Se proporciona también los valores para el conjunto de la población.

Existe luego un grupo de opciones referidas a "Modelos de superpoblación". En este caso se puede determinar el número de veces que se va a repetir el proceso de comparación de métodos de selección aplicado a una determinada estructura poblacional. En el caso más sencillo, que no es propiamente un modelo de superpoblación, definimos una única población finita, que coincide

con la población natural, y obtenemos para los métodos especificados, el error debido al muestreo para cada método. En el caso de utilizar dos o más poblaciones finitas entonces el programa requiere que la población sobre la que se apliquen los métodos de selección sea aleatoria y no una población natural. Un número suficientemente elevado de poblaciones finitas para obtener resultados estables puede estar entre 40 y 50, no obstante es posible aumentar este número en función de los datos que se estén manejando. Por simplicidad y para poder comprender mejor las comparaciones se han formado tres grandes bloques: así es posible obtener la esperanza de la varianza del estimador y su correspondiente varianza respecto del modelo; en un segundo grupo de opción es posible obtener cálculos similares a los anteriores pero referidos a la esperanza y varianza respecto del modelo del estimador de la varianza; el tercer grupo hace referencia, en la misma línea anterior, a la esperanza y varianza respecto del modelo de la varianza del estimador de la varianza.

La opción de "Coeficiente de correlación" nos proporciona, para un determinado método el valor del coeficiente de correlación intraclásico y ciertos parámetros relacionados con esta medida. La opción "Correlograma" obtiene para cada uno de los métodos la representación gráfica del valor del coeficiente anterior para diferentes tamaños de muestra.

La "Descomposición de la varianza" permite obtener el cuadro de análisis de la varianza distinguiendo según si la fuente de variación es entre o dentro de muestras.

Por último, en este apartado de cálculos, la opción de "Estimación "Bootstrap", considera que la población que se está analizando es una muestra. El procedimiento permite definir al usuario el número de veces que se van a repetir cada observación para auto-generar la población. Después obtiene una representación gráfica del espacio muestral sin reposición y calcula las estimaciones "Bootstrap". En cuanto a la opción "Estimación "Jackknife"" considera también la población como una muestra y obtiene muestras de tamaño n-1 sin reposición para estimar la varianza de la característica que se esté considerando.

3.2. PRINCIPALES LISTADOS OBTENIDOS CON POSDEM

Los procedimientos de cálculo obtienen salidas impresas tanto a tabla de celdas, como a informe. Además es posible obtener otro tipo de listados bajo la opción de "Listados", desde esta opción podemos obtener la "Ficha técnica" de la Encuesta, este listado nos proporciona información sobre las principales características del diseño que se está realizando: El error de muestreo, el método elegido, las características de la población marco, el tamaño de muestra y otras indicaciones sobre intervalos de confianza.

Es posible obtener también distintos listados de las "Unidades de cada muestra", así obtendremos un listado de cada muestra con información sobre el valor de la variable para esa unidad, el valor de la variable auxiliar, o el identificador de cada unidad, estas últimas siempre que se hayan definido en el editor de datos.

En cuanto a la opción de "Eficacia relativa" el programa permite comparar el método que se esté analizando con el muestreo sin reposición con probabilidades iguales y con el muestreo estratificado con una unidad por estrato. En este apartado tendremos distintas posibilidades: por una parte podremos estudiar la eficacia relativa de estos métodos por el cociente entre sus varianzas; por otra

los tamaños de muestra que hubiesen sido necesarios para diferentes errores de muestreo en el caso de muestreo sin reposición. Y, por último que tamaño de muestra hubiese sido necesario para con una determinada varianza en los casos de sin reposición y estratificado con una unidad.

Las opciones restantes de este apartado permiten llevar a cabo "Gestión del fichero histórico" , esta gestión consiste básicamente, en que el programa da la opción de guardar los resultados de diseños alternativos para compararlos entre sí. Estos resultados pueden observarse en forma de tabla o de gráfico. Por último la opción "Obtención de una muestra" permite varias alternativas por una parte obtener los identificadores que dan a conocer cuáles son las unidades de la población que deberán ser efectivamente investigadas en la encuesta por muestreo y por otra obtener un fichero en dbase con los valores de identificación, de la variable de estudio y de la variable auxiliar. Fichero que puede utilizarse como entrada en nuevos experimentos fundamentalmente para realizar estimaciones de la varianza, al considerar la muestra como una población. Por último en esta opción esta disponible la alternativa "Obtención de una población" que permite guardar los valores de una determinada población.

3.3. GRÁFICOS PREDEFINIDOS EN LA APLICACIÓN

Cualquiera de las salidas anteriores es susceptible de exportación a aplicaciones gráficas que permiten ver los resultados desde una óptica gráfica. POSDEM incorpora, dentro de la misma aplicación una opción con gráficos predeterminados. Dentro del apartado de "Gráficos" se han considerado ocho grupos y un total de diecisiete posibilidades diferentes. Por este motivo y debido a que la mayor parte de los gráficos tienen un equivalente resultado de las opciones de cálculo o listado nos vamos a limitar a explicar sólo algún gráfico especial.

Así, tendremos el gráfico de "Estimaciones y errores", en el cual en ordenadas se representan los valores de la variable: su media y los límites de confianza; en abcisas se representa el resultado de cada reiteración. Cada punto dentro del gráfico representa el valor obtenido para el estimador media o proporción y se representa mediante una circunferencia. El radio de la circunferencia se traza proporcionalmente al valor del error de muestreo. Entonces, la circunferencia será grande si en una determinada muestra el error de muestreo asociado al estimador es superior a la media de todos los errores de muestreo. Si la pantalla del ordenador en el que se ejecuta la aplicación es de color, entonces la circunferencia se representará en rojo en el caso de un error superior a la media; y a la inversa

estimaciones con una error de muestreo pequeño, se representarán con un punto o circunferencia pequeño y resaltado en verde. En este texto sólo será visible en el gráfico el tamaño de la circunferencia.

Otro gráfico especialmente interesante es el que hemos denominado "Supergrama" que consiste en representar en abcisas diferentes poblaciones que con la misma estructura se generan de forma aleatoria, en ordenadas se representa la varianza del estimador para cada una de esas poblaciones según diferentes métodos. Existe un primer grupo para métodos con probabilidades iguales, otro para probabilidades proporcionales y, al fin otro que combina los dos procedimientos probabilísticos.

El "Correlograma" permite obtener una representación gráfica del coeficiente de correlación intraclásica para diferentes tamaños de muestra.

La representación, para un determinado tamaño de muestra, del "Coeficiente de correlación intraclásico" se ha puesto en relación con los cocientes 1/n y 1/(n-1) dado que entre esos límites es posible obtener una estimación muy acurada del error de muestreo en diseños sistemáticos.

Otros gráficos muestran distribuciones de frecuencias y alternativas a los gráficos comentados anteriormente.

3.4. OPCIONES DEL USUARIO

El apartado de "Opciones" permite al usuario determinar ciertas características que el programa deberá tener en cuenta al ejecutar otras posibilidades del menú. Se trata de especificaciones de configuración. Al elegir esta opción podremos determinar el número de decimales que queremos utilizar en las salidas impresas. Tendremos también la posibilidad de determinar el conjunto de los métodos que se utilizaran en las comparaciones de superpoblación. Será posible también determinar otros aspectos, entre los que destacan: en primer lugar tendremos la opción de "Selección circular" , si se elige aparecerá una señal de selección ("v") en su margen izquierdo. Esto significa que al utilizar métodos sistemáticos de muestreo se utilizará una definición circular del procedimiento con lo cual no será necesario que la población sea divisible de forma exacta entre el tamaño de muestra para determinar el número de clases con un entero. Por defecto la aplicación considera esta opción desactivada. Las siguientes tres opciones que empiezan con "Orden aleatorio" hacen referencia al supuesto que se utiliza para calcular la varianza del estimador en cada muestra. Las otras dos opciones dentro de este grupo son: "Técnica del lazo" que une dos unidades consecutivas simulando un mismo estrato, y "Semimuestras" que utiliza un procedimiento de cálculo para estimar el error basado en obtener un número suficiente de muestras de cada muestra y

casarlas entre sí por pares de forma que es posible obtener estimaciones del error. Otra opción, "Varianza intraconglomerados" aplicable al procedimiento de dos conglomerados consiste en utilizar la formula en función de la varianza entre conglomerados. Por otra parte podemos definir al programa el tipo de datos que deberá procesar, con la opción "Tipo de variable", será posible determinar si los cálculos deberán realizarse sobre variables cuantitativas o cualitativas. También podremos indica si nos interesa trabajar con estimadores de la media, del total o estimadores mejorados de regresión.

3.5. VENTANAS Y AYUDA EN LÍNEA

La opción de "Ventana" es una opción normalizada en programas de tipo Windows permite organizar el aspecto de las ventanas abiertas por la aplicación. Permite también moverse a través de ellas. Básicamente las ventanas de la aplicación POSDEM son las siguientes:

1.- Pantalla general de la aplicación: contiene a todas las demás, es la que aparece al iniciar la aplicación.

2.- Listado de resultados: es una pantalla de modo texto donde la aplicación deja explicaciones e informes dependiendo de los pasos que se ejecuten.

3.- Tabla de resultados: similar a la anterior, pero deja los resultados en forma de tabla numérica y no en formato informe escrito.

4.- Gráficos: es una pantalla que recoge los gráficos que el usuario va demandando al sistema.

5.- Generador de estructuras poblacionales: permite definir fórmulas matemáticas de poblaciones naturales.

6.- Editor de datos: permite seleccionar variables, crear nuevas y editar, ordenar o suprimir datos que configuran la población.

7.- Generador de componentes principales: esta pantalla contiene el modulo de las componentes principales.

8.- Pantalla de ayuda: contiene las posibilidades de información que el usuario precisa para hacer funcionar correctamente el programa.

La opción "Ayuda" permite acceder a un sistema hipertexto de ayuda en línea. Las posibilidades de esta ayuda están normalizadas en Windows y permiten obtener información, mediante índices, mediante búsquedas o mediante glosarios, de las posibilidades del programa.

CAPÍTULO 4

MUESTREO ALEATORIO SIMPLE

4.1. MUESTREO CON REPOSICIÓN

La teoría del muestreo en poblaciones finitas tiene como objetivo obtener métodos de selección y estimación de forma que sea posible sustituir toda la información que suministra una población por la que suministra una muestra. Para proceder de esta forma se dan razones de coste, de calidad de los datos y de imposibilidad, en algunos casos, de analizar la población completa. Los métodos de muestreo se diferencian unos de otros en el tipo de información que utilizan y en la forma de utilizarla. En los próximos capítulos vamos a desarrollar ejemplos de las técnicas de selección de muestras más habituales. Únicamente se va a aportar una modificación al método de selección sistemática y al método de conglomerados.

Para determinar qué método se va a utilizar, es necesario analizar toda la información disponible sobre el fenómeno que se quiere investigar. El conjunto de esta información es lo que se denomina en un sentido amplio marco de la encuesta. En un sentido más restringido el marco consiste en un listado de unidades que tienen, al menos, la siguiente información: Un número de identificación y el valor de una variable. Esta variable se supone significativa para el diseño.

En este capítulo, y con el fin de mostrar los procedimientos de selección del muestreo aleatorio con y sin reposición, se va a

utilizar para resolver ejemplos poblaciones del tipo $X_i = i$ para $i = 1,2,3...N$. Este tipo de ejemplos es bastante frecuente en textos de muestreo que, con fines didácticos, utilizan, por ejemplo, una población de tamaño 3 y muestras de tamaño 2. En los siguientes apartados se realizan algunos ejemplos que permiten comprender los mecanismos en los que se basan ciertas técnicas de la teoría del muestreo en poblaciones finitas y la aplicación informática POSDEM.

Dado un esquema con un tamaño de muestra (n), selección con reposición y donde las muestras que constan de las mismas unidades en distinto orden se consideran idénticas, el espacio muestral está compuesto por las combinaciones con repetición de N elementos tomados de n en n.

Así la probabilidad de seleccionar una determinada muestra, $(u_1, u_2, \ldots, u_n)$, es:

$$P(u_1, u_2, \ldots, u_n) = \frac{1}{\binom{N+n-1}{n}}$$

Donde N representa el tamaño de la población y n es el tamaño de la muestra.

Si el esquema es sin reposición, entonces el espacio muestral es: $S(\underline{x}) = \binom{N}{n}$

esto es, el conjunto de todas las muestras posibles está formado por combinaciones de N elementos tomados de n en n, y la probabilidad de seleccionar una determinada muestra será:

$$P(u_1, u_2, \ldots, u_n) = \frac{1}{\binom{N}{n}}$$

Para el caso de muestreo con reposición y datos cuantitativos, las fórmulas para el estimador media, la varianza del estimador y su estimador vendrán dadas respectivamente por:

$\hat{\bar{x}} = \frac{\sum_{i=1}^{n} x_i}{n}$ $V(\hat{\bar{x}}) = \frac{\sigma^2}{n}$ $\hat{V}(\hat{\bar{x}}) = \frac{\hat{S}^2_{n-1}}{n}$ Con reposición se verifica que $E(\hat{S}^2_{n-1}) = \sigma^2$ y los límites de confianza para el 95% vienen dadas por $\hat{\bar{x}} + -2 * \sqrt{\hat{V}(\hat{\bar{x}})}$

Hay que resaltar que para muestreo sin reposición $E(\hat{S}^2_{n-1}) = S^2_{n-1}$ y por tanto $V(\hat{\bar{x}}) = (1-f)\frac{S^2_{n-1}}{n}$ tiene un estimador insesgado en

$$\hat{V}(\hat{\bar{x}}) = (1-f)\frac{\hat{S}^2_{n-1}}{n}$$

4.1.1. EJEMPLO RESUELTO CON LÁPIZ Y PAPEL

Si tenemos una población de tres individuos en los que se realiza la medición de un valor variable que se quiere estimar, tendremos:

Unidades población u_i = 1 2 3

Valor de la variable (X_i) = 1 2 3

Si se seleccionan las muestras con reposición y probabilidades iguales, el espacio muestral está formado por los siguientes pares de observaciones posibles:

$$S(\underline{x}) = \begin{bmatrix} (1,1) & (1,2) & (1,3) \\ (2,1) & (2,2) & (2,3) \\ (3,1) & (3,2) & (3,3) \end{bmatrix}$$

La función de probabilidad asociada a este método proporciona la siguiente probabilidad para cada muestra:

$$P(\underline{x}) = \begin{bmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{bmatrix}$$

La función del estimador media muestral asigna a cada muestra un valor estimado

$$\hat{\bar{x}} = \begin{bmatrix} 1 & 1.5 & 2 \\ 1.5 & 2 & 2.5 \\ 2 & 2.5 & 3 \end{bmatrix}$$

El valor esperado del estimador es la suma de valores por sus probabilidades

$$E(\hat{\bar{x}}) = 1*\frac{1}{9} + 1.5*\frac{1}{9} + \ldots + 3*\frac{1}{9} = 2$$

Donde se puede observar que se trata de un estimador insesgado puesto que su valor esperado coincide con el valor poblacional que se quiere estimar.

En cuanto a la varianza del estimador podemos asimismo utilizar en el ejemplo la distribución en el muestreo completa[1]; obteniendo:

$$(1) \quad V(\hat{\bar{x}}) = \frac{1}{3}$$

expresión, que si se conoce la varianza poblacional, y dado que para este método de muestreo existe un estimador insesgado, tendremos:

[1] La aplicación POSDEM utilizará la formación del espacio muestral completo para calcular la varianza del estimador en el caso del muestreo sistemático. Cuando el método de selección sea distinto se utilizará una representación de ese espacio, basado en reiteraciones, para evitar la dificultad que puede suponer un número muy elevado de muestras posibles.

$$(2) \quad V(\bar{x}) = \frac{\sigma^2}{n} = \frac{0,8164^2}{2} = \frac{1}{3}$$

En la realización de una encuesta no se conoce la varianza pero se dispone, si se utiliza selección con reposición, de un estimador insesgado con la cuasivarianza muestral. En este caso la varianza del estimador será a su vez un estimador al depender de los valores de cada muestra. De esta forma, para cada muestra existirá un estimador de la varianza del estimador, con su función de probabilidad asociada.

$$\hat{V}(\bar{x}) = \begin{bmatrix} 0 & 0.25 & 1 \\ 0.25 & 0 & 0.25 \\ 1 & 0.25 & 0 \end{bmatrix}$$

Con su correspondiente valor esperado

$$E[\hat{V}(\bar{x})] = \frac{1}{9}[0.25 + 1 + 0.25 + 0.25 + 1 + 0.25] = \frac{3}{9} = \frac{1}{3}$$

y su correspondiente varianza. La varianza del estimador de la varianza mide la estabilidad de la varianza estimada.

$$V[\hat{V}(\bar{x})] = \frac{1}{9}[4*(0.25-0.33)^2 + 2*(1-0.33)^2 + 3*(0-0.33)^2] = 0.1381$$

La importancia de que la varianza del estimador de la varianza este bajo control es evidente. Si de seleccionar una muestra a seleccionar otra, el error de muestreo va a ser muy diferente, se

perderá la confianza en la evaluación del estimador utilizando el concepto de error de muestreo.

El objetivo de este apartado ha sido introducir los tres conceptos que tendrán un papel destacado en el resto del trabajo: varianza del estimador; esperanza del estimador de la varianza; y varianza del estimador de la varianza. Cada uno de estos parámetros es susceptible de tratamiento bajo la óptica de los modelos de superpoblación, dando lugar a los conceptos de esperanza y varianza respecto del modelo de cada uno de los tres conceptos.

4.1.2. EJEMPLO RESUELTO CON HOJA DE CÁLCULO

Con el único fin de mostrar cuál es la estructura de la aplicación POSDEM vamos a desarrollar el mismo ejemplo anterior, pero con la ayuda de una hoja de cálculo. Los resultados serán los mismos que los obtenidos hasta ahora, si bien se han ampliado a los momentos y al resto de los parámetros, dadas la facilidad de cálculo. Con la aplicación POSDEM es posible obtener los mismos resultados.

Tabla 8.- Estimaciones para el espacio muestral con reposición.

MUESTRAS	MED[2]	VAR[3]	DES[4]	Li[5]	Ls	CVS	SG2[6]	SG	CU[7]	CD	CVS	M2[8]	M3	M4
1 2	1,50	0,25	0,50	0,5	2,5	33%	0,25	0,50	0,50	0,71	47%	2,50	4,50	8,50
1 3	2,00	1,00	1,00	0	4	50%	1,00	1,00	2,00	1,41	71%	5,00	14,0	41,00
2 1	1,50	0,25	0,50	0,5	2,5	33%	0,25	0,50	0,50	0,71	47%	2,50	4,50	8,50
2 3	2,50	0,25	0,50	1,5	3,5	20%	0,25	0,50	0,50	0,71	28%	6,50	17,5	48,50
3 1	2,00	1,00	1,00	0	4	50%	1,00	1,00	2,00	1,41	71%	5,00	14,0	41,00
3 2	2,50	0,25	0,50	1,5	3,5	20%	0,25	0,50	0,50	0,71	28%	6,50	17,5	48,50
1 1	1,00	0,00	0,00	1	1	0%	0,00	0,00	0,00	0,00	0%	1,00	1,00	1,00
2 2	2,00	0,00	0,00	2	2	0%	0,00	0,00	0,00	0,00	0%	4,00	8,00	16,00
3 3	3,00	0,00	0,00	3	3	0%	0,00	0,00	0,00	0,00	0%	9,00	27,0	81,00
MEDIA[9]	2,00	0,33	0,44	1,11	2,89	23%	0,33	0,44	0,67	0,63	32%	4,67	12,0	32,67
VARIANZA	0,33	0,14	0,14	0,88	0,88	4%	0,14	0,14	0,56	0,27	7%	5,44	60,3	602,7
DESVIACION	0,58	0,37	0,37	0,94	0,94	19%	0,37	0,37	0,75	0,52	27%	2,33	7,77	24,55
COEF.VAR	29%	112%	83%	84%	32%	83%	112	83%	112%	83%	83%	50%	65%	75%
LIM.SUP	3,15	1,08	1,18	2,98	4,76	61%	1,08	1,18	2,16	1,67	86%	9,33	27,5	81,77
LIM.INF	0,85	-0,41	-0,29	-	1,02	-15%	-	-	-0,82	-	-	0,00	-	-16,44
POBLA_CR[10]	2,00	0,33	0,58	0,76	0,87	29%	0,67	0,82	1,00	1,00	50%	4,67	12,0	32,67

[2] MED es el estimador de la media calculado para cada muestra.
[3] VAR es el estimador de la varianza calculado para cada muestra
[4] DES es el error debido al muestreo. Raíz de VAR.
[5] Li y Ls son los límites de confianza la 95% supuesto normalidad.
[6] SG2 es la varianza poblacional
[7] CU es la cuasivarianza y Cd cuasidesviación.
[8] M2-4 momentos respecto al origen.
[9] Parámetros calculados sobre el conjunto de las muestras posibles.
[10] Parámetros calculados sobre el conjunto de la población utilizando muestreo con reposición.

En este cuadro se puede observar en las dos primeras columnas las muestras posibles. Esta parte tiene su reflejo en la aplicación POSDEM en el listado de unidades muestrales. Después, en filas aparece en primer lugar los cálculos para estimar la media, la varianza de la media, su desviación, los límites de confianza y el coeficiente de variación. Con POSDEM estos datos se obtienen con el listado de cálculos para cada muestra. En un segundo apartado en la misma fila están los cálculos necesarios para varianza muestral, la cuasivarianza, sus desviaciones, el coeficiente de variación muestral y los momentos de segundo, tercer y cuarto orden.

En este cuadro puede destacarse un segundo bloque con la información correspondiente a todas las muestras posibles. En la primera fila aparece la media, tanto de la misma media como de cada parámetro que se estima. Se puede observar como la media de todas las varianzas de la media coincide con su valor poblacional (0.33). Sin embargo la desviación y el coeficiente de varlación no son estimadores insesgados. En la aplicación se han sustituido los valores correspondientes de la desviación por la raíz de la media de las varianzas, para disponer así de un valor insesgado. La segunda fila muestra la varianza de cada uno de los parámetros. Así tenemos la varianza de la media, la varianza de la varianza etc ... El resto de las filas se explica por si mismo. Este listado en la aplicación es el listado de esperanzas y varianzas para todas las muestras.

4.1.3. EJEMPLO RESUELTO UTILIZANDO LA APLICACIÓN POSDEM

De los métodos de selección implementados en la aplicación POSDEM, a excepción de los métodos sistemáticos para los cuales se obtiene todo el espacio muestral para los restantes el espacio muestral se representa mediante un conjunto de reiteraciones que pueden oscilar de un mínimo de 30 a un máximo que depende de la memoria del ordenador. Para estos métodos , básicamente muestreo aleatorio simple con y sin reposición, estratificado con una unidad, probabilidades desiguales y muestreo bietápico hay que resaltar que utilizando este procedimiento computarizado de cálculo los resultados no son tan precisos debido a la dificultad de obtener todas las muestras posibles. Por tanto, en lugar del espacio muestral completo se obtienen reiteraciones que simulan la formación del espacio muestral y que proporcionan un resultado suficientemente acurado.

Se reproducen a continuación listados más significativos para este ejemplo. En el informe, elaborado por el programa con salida a un documento "Listados de POSDEM" se proporcionan una serie de notas y sugerencias al usuario. Los listados con cálculos se han importado a través de portapapeles de "Windows" al informe de "Listado". Así en el listado correspondiente a los cálculos para cada muestra se pueden observar los resultados obtenidos para las veinte muestras seleccionadas aleatoriamente con reposición. Siguiendo la sugerencia del informe, y con los mismos datos, seleccionamos en el menú de listados la opción de

listado de cálculos para el conjunto de las muestras y podremos observar la matriz de datos correspondientes a las esperanzas y varianzas para todas las muestras.

POSDEM

Ficha Técnica de la Encuesta: Datos del diseño

Población marco utilizada : Ficticia
Variable de estudio : 1,2,3
Variable auxiliar : No

Método de muestreo empleado : CON REPOSICIÓN
Tamaño de población = 3
Tamaño de muestra = 2

Media poblacional = 2
Desviación típica (n-1) = 1

Fracción de muestreo = 2/3
Desviación típica del estimador = ,58
Error de muestreo CV% (Basado en una representación de todas las muestras) = 29,38%

POSDEM

Estimadores para todas las muestras: muestreo con reposición

Notas:

1.- En filas se representa, la media, varianza, ... calculadas para todas las muestras obtenidas. En columnas figura el estimador al que se refieren los cálculos. Ejemplo, el valor media, varmed representa la media de todas las varianzas del estimador media, calculada en base a todas las muestras.

2.- Estimadores sesgados de Desmed y Cvsmed, corregidos. En la línea correspondiente figuran los valores de Desmed y de Cvsmed calculados como raíz de Varmed. El resto de los cálculos, por ejemplo varianza de S se han mantenido en su formato original

	Media	Varmed	Desmed	li	ls	Cvsmed%
MEDIA	1,98	,34	,58	,81	3,14	29,42
VARIANZA...	,34	,12	,11	,99	,58	311,00
DESVIACION	,58	,35	,33	,99	,76	17,64
CO.VAR .(%)	29,39	102,84	70,42	96,92	26,08	66,97
LIM.SU....	3,14	1,03	1,14	3,01	4,45	61,60

LIM.IF.... ,81 -,36 -,19 -,96 1,40 -8,94

POSDEM

Estimadores para cada muestra: muestreo con reposición

Notas:

1.- En filas se representa cada muestra obtenida, en columnas los valores estimados.

2.- Ejemplo: el valor Muestra(4),Varmed representa la varianza del estimador media calculado para la muestra 4.

3.- Muestras obtenidas = 20 Tamaño de muestra = 2 Tamaño de población = 3

Sugerencia:

1.- Compruebe como los estimadores de cada muestra oscilan entre los intervalos de
confianza calculados para el conjunto de todas las muestras. Para esto seleccione primero la opción cálculos para todas las muestras, posteriormente seleccione la opción cálculos para cada muestra.

	Media	Varmed	Desmed	li	ls	Cvsmed%
Muestra(1)	2	0	0	2	2	0
Muestra(2)	1,5	,25	,5	,5	2,5	33,33
Muestra(3)	2	1	1	0	4	50
Muestra(4)	1,5	,25	,5	,5	2,5	33,33
Muestra(5)	1,5	,25	,5	,5	2,5	33,33
Muestra(6)	2	1	1	0	4	50
Muestra(7)	2,5	,25	,5	1,5	3,5	20
Muestra(8)	2	1	1	0	4	50
Muestra(9)	1	0	0	1	1	0
Muestra(10)	1,5	,25	,5	,5	2,5	33,33
Muestra(11)	1,5	,25	,5	,5	2,5	33,33
Muestra(12)	2	1	1	0	4	50
Muestra(13)	1,5	,25	,5	,5	2,5	33,33
Muestra(14)	3	0	0	0	0	0
Muestra(15)	1,5	,25	,5	,5	2,5	33,33
Muestra(16)	3	0	0	0	0	0
Muestra(17)	3	0	0	0	0	0
Muestra(18)	1,5	,25	,5	,5	2,5	33,33
Muestra(19)	2,5	,25	,5	1,5	3,5	20
Muestra(20)	2,5	,25	,5	1,5	3,5	20

En este informe se puede apreciar como la media, de la varianza del estimador, lo que se denomina en la aplicación el cuadrado del error de muestreo calculado con la información muestral, no coincide exactamente con la varianza de la media, el error de muestreo calculado con las reiteraciones, y ambas con el verdadero valor calculado al conocer la varianza de la población (0,33). Esto es debido, en este caso, a que la aplicación no obtiene todas las muestras posibles sino únicamente un conjunto de reiteraciones que representan a ese espacio muestral. Para el caso de muestreo sistemático, en sus distintas modalidades, las reiteraciones calculan exactamente el espacio muestral. Sin embargo en estos casos las estimaciones referidas no coinciden por el problema que se presenta al estimar el error con este método. En un próximo capítulo se dedicará un apartado a comprobar como las estimaciones, para el caso sin reposición, se estabilizan alrededor del verdadero valor a medida que las reiteraciones sobrepasan el número de treinta o cuarenta.

4.1.4. EJEMPLO PARA MUESTREO CON REPOSICIÓN Y DATOS CUALITATIVOS

Con una población de tres unidades $u_i = 1\ 2\ 3$

Valores de la variable de estudio para cada $u_i = 0\ 1\ 1$

Donde la proporción poblacional viene dada por $P = 2/3$ y su varianza poblacional por $\sigma^2 = 2/3\ (1 - 2/3) = 2/9$

La interpretación de estos datos es, que la primera unidad no tiene la característica que se está estudiando, por ejemplo, el paro; mientras que la segunda y la tercera unidad si la tienen.

El espacio muestral y la distribución del estimador de la proporción vendrán dados por:

$$S(\underline{x}) = \begin{bmatrix} 0.0 & 0.1 & 0.1 \\ 1.0 & 1.1 & 1.1 \\ 1.0 & 1.1 & 1.1 \end{bmatrix}$$

Este es el conjunto de las muestras posibles con la representación del valor de la variable. Los valores que puede tomar el estimador de la proporción son, según la muestra que se haya seleccionado:

$$\hat{P} = \begin{bmatrix} 0 & .5 & 0.5 \\ 0.5 & 1 & 1 \\ .5 & 1 & 1 \end{bmatrix}$$

El valor esperado del estimador será el valor poblacional al tratarse de un estimador insesgado.

$$E(\hat{P}) = \frac{1}{9}[.5 + .5 + .5 + .5 + 4] = \frac{6}{9} = \frac{2}{3}$$

La varianza del estimador proporción será , utilizando la información de todas las muestras

$$(1) \quad V(\hat{P}) = \frac{1}{9}[(.5-\frac{2}{3})^2 * 4 + (1-\frac{2}{3})^2 * 4 + (0-\frac{2}{3})^2] = 0.11$$

Como en este caso se conoce la varianza poblacional se puede calcular

$$(2) \quad V(\hat{P}) = \frac{PQ}{n} = \frac{\frac{2}{9}}{2} = .11$$

Que es la varianza del estimador de la proporción para las especificaciones de este problema. Con la aplicación POSDEM se puede seguir también este tipo de ejemplos, con la ventaja de cálculo que proporciona el ordenador en cuanto a tamaños de población y diversidad de métodos de selección aplicables.

4.2. MUESTREO SIN REPOSICIÓN

Con este tipo de muestreo las unidades que forman la población únicamente pueden pertenecer a la muestra una vez. Se trata de seleccionar las unidades que van a formar parte de la muestra de manera que si una unidad pertenece a la muestra ya no podrá volver a ser seleccionado. Para verlo más gráficamente, la regla práctica consiste en introducir en una urna tantas bolas como unidades existan en la población, para después seleccionar una primera unidad que pasará a formar parte de la muestra. La bola correspondiente a esta unidad no vuelve a formar parte de la urna, por tanto esa unidad tampoco puede pertenecer de nuevo a la muestra. El espacio muestral estará compuesto por las combinaciones de N elementos tomados de n en n.

La probabilidad de una determinada muestra será:

$$P[u_1, u_2, \ldots, u_n] = \frac{1}{\binom{N}{n}}$$

Las fórmulas para estimar la media, el total o la proporción son iguales al caso con reposición. En cuanto a la estimación de la varianza del estimador, esta se diferencia en el factor corrector de poblaciones finitas.

$$\hat{V}(\hat{\bar{x}}_{sr}) = \frac{N-n}{N}\hat{V}(\hat{\bar{x}}_{cr})$$

De esta forma, el muestreo sin reposición será más preciso que el muestreo con reposición dependiendo de la fracción de muestreo. Si el tamaño de muestra, en relación con el tamaño de población proporciona una fracción de muestreo muy pequeña, entonces el factor de corrección de poblaciones finitas es despreciable, y en la práctica, es indiferente aplicar un método u otro.

Sin embargo, si la fracción de muestreo es significativa la reducción en la varianza del estimador es significativa también. En este último caso es conveniente aplicar el esquema sin reposición.

Por ejemplo, si la población está compuesta por 400 individuos y se selecciona una muestra de tamaño 80, la fracción de muestreo será igual al 20 %, esto supone un factor de corrección (1-0,2)=0,8, esto es, en este caso como consecuencia de utilizar un esquema sin reposición se producirá una disminución del 20% en el error debido al muestreo.

4.2.1. EJEMPLO DE MUESTREO SIN REPOSICIÓN

Con una población u_i = 1 2 3 y muestras de tamaño 2, se puede construir todo el espacio muestral y analizar, de forma similar, a como se hizo en el ejemplo anterior el error debido al muestreo.

Tabla 9.- Estimaciones para el espacio muestral sin reposición.

Muestras	MED[11]	VAR[12]	DES[13]	LI[14]	Ls	CVS	SG2[15]	SG	CU[16]	CD	CVS	M2[17]	M3	M4
1 2	1,50	0,08	0,29	0,92	2,08	19%	0,25	0,50	0,50	0,71	47%	2,50	4,50	8,50
1 3	2,00	0,33	0,58	0,85	3,15	29%	1,00	1,00	2,00	1,41	71%	5,00	14,00	41,00
2 1	1,50	0,08	0,29	0,92	2,08	19%	0,25	0,50	0,50	0,71	47%	2,50	4,50	8,50
2 3	2,50	0,08	0,29	1,92	3,08	12%	0,25	0,50	0,50	0,71	28%	6,50	17,50	48,50
3 1	2,00	0,33	0,58	0,85	3,15	29%	1,00	1,00	2,00	1,41	71%	5,00	14,00	41,00
3 2	2,50	0,08	0,29	1,92	3,08	12%	0,25	0,50	0,50	0,71	28%	6,50	17,50	48,50
MEDIA[18]	2,00	0,17	0,38	1,23	2,77	20%	0,50	0,67	1,00	0,94	0,49	4,67	12,00	32,67
VARIANZA	0,17	0,01	0,02	0,24	0,24	0%	0,13	0,06	0,50	0,11	0,03	2,72	30,17	301,39
DESVIACI	0,41	0,12	0,14	0,49	0,49	7%	0,35	0,24	0,71	0,33	17%	1,65	5,49	17,36
COEF.VAR	20%	71%	35%	40%	18%	36%	71%	35%	71%	35%	36%	35%	46%	53%
LIM.SUP	2,82	0,40	0,66	2,21	3,75	34%	1,21	1,14	2,41	1,61	83%	7,97	22,98	67,39
LIM.INF.	1,18	-0,07	0,11	0,25	1,79	6%	-0,21	0,20	-0,41	0,28	14%	1,37	1,02	-2,05
POBLACIÓN[19]	2,00	0,17	0,41	0,64	0,80	20%	0,67	0,82	1,00	1,00	50%	4,67	12,00	32,67

[11] MED es el estimador de la media calculado para cada muestra.
[12] VAR es el estimador de la varianza calculado para cada muestra
[13] DES es el error debido al muestreo. Raíz de VAR.
[14] Li y Ls son los límites de confianza la 95% supuesto normalidad.
[15] SG2 es la varianza poblacional
[16] CU es la cuasivarianza y Cd cuasidesviación.
[17] M2-4 momentos respecto al origen.
[18] Parámetros calculados sobre el conjunto de las muestras posibles.
[19] Parámetros calculados sobre el conjunto de la población para muestreo sin reposición.

Los principales resultados son similares en su interpretación al ejemplo anterior. Ahora bien existen algunos aspectos que es necesario resaltar. La varianza del estimador ha disminuido, pasando de 0,33 a 0,17. Sin embargo la reducción más notable ha sido en el error del error, esto es ha disminuido la varianza de la varianza del estimador, y por tanto ha aumentado la estabilidad de la varianza. Esta estimación ha pasado de 0,14 a 0,01. Esto quiere decir que al seleccionar diferentes muestras el error debido al muestreo no sufrirá variaciones excesivamente grandes, en comparación con el método de reposición.

4.3. PRÁCTICAS PARA EL MUESTREO ALEATORIO SIMPLE Y POSDEM

Estas practicas pueden realizarse utilizando la aplicación POSDEM.

Vamos a considerar una población donde el número de identificación de cada unidad coincide con la característica que se pretende medir.

Unidades poblacionales 1 2 3 4 5 6 7 8 9 10

Valores de la variable 1 2 3 4 5 6 7 8 9 10

Los parámetros poblacionales son:

$$Media = \overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = 5.5$$

$$Varianza = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = (2.87)^2 = 8.23$$

1.-) Con un procedimiento de muestreo con reposición y tamaño de muestra igual a 4.

a) Determinar las muestras posibles.

b) Calcular para cada muestra su media, la varianza de la media, desviación, límites de confianza y coeficiente de variación.

c) Utilizando como base cinco muestras obtenidas calcular los valores esperados del estimador de la media y de todos los demás parámetros calculados en el apartado anterior.

d) Representar gráficamente las muestras obtenidas, la media de cada una de ellas y la importancia relativa del error de muestreo asociado a cada estimador.

2.-) Con los mismos datos calcular el tamaño de muestra para el caso de sin reposición y errores de muestreo relativos del 2%, 5%, 10% y 15%.

3.-) Grabar los datos del problema uno y repetir la misma práctica descrita en los ejemplos uno y dos para el caso de muestreo sin reposición. Grabar los datos en el fichero histórico y realizar un análisis comparativo de ambos métodos.

4.-) Repetir las prácticas 1, 2 y 3 para una población de tamaño 100, muestras de tamaño 10 y número de reiteraciones 40.

CAPÍTULO 5

MUESTREO SISTEMÁTICO I

5.1. MUESTREO SISTEMÁTICO CON INTERVALO CONSTANTE

El muestreo sistemático tiene como objetivo obtener con un método sencillo de aplicación un efecto similar al obtenido con la estratificación y extender la muestra a toda la población. Para que este método sea efectivo necesita que los elementos de la población se puedan ordenar con un criterio relevante para la investigación y no introducir regularidades ocultas.

Con esta técnica se obtienen tantos estratos como elementos se quieran incorporar a la muestra. La muestra estará formada en este caso con un elemento de cada estrato. Esta muestra se puede considerar como un conglomerado.

Este método consiste, en una vez ordenadas las unidades de la población, dividir la población en k grupos iguales, de forma que N/n = k. Después, seleccionar un elemento del primer grupo y los sucesivos elementos que ocupen la misma posición.

Por ejemplo, dada una población donde los valores de la variable en cada unidad coincide con el valor que las identifica.

$$Xi = i \text{ con } i = 1,2, \ldots 12$$

Se puede considerar, con fines didácticos, que cada unidad representa una sección de un determinado estrato. Y que el valor de la variable medida en cada unidad representa el número de personas que

en el momento del censo se clasificaron en paro. Las unidades están ordenadas conforme el valor de la variable. Si fijamos un tamaño de muestra de 4 unidades, se procederá de la siguiente forma:

Para dividir la población en grupos se obtiene el valor k= 12/4 = 3 período que se utiliza para dividir la población en n=4 grupos iguales. Esto es, el número de unidades de la población dividido por el número de unidades de la muestra.

$$k = 3 \quad n = 4 \quad N = 12$$

Se selecciona en el primer grupo uno de los 3 elementos con probabilidad igual a

$$\frac{1}{k} = \frac{1}{3}$$

Si por ejemplo se hubiese seleccionado el valor 2, éste condicionará el resto de la muestra, que estará configurado por aquellos elementos que ocupan en los otros grupos la misma posición relativa. Esto es, la muestra quedaría compuesta por los siguientes elementos (2, 5, 8, 11).

Así se ha seleccionado la unidad U_2, U_{2+K}, U_{2+2K}, U_{2+3K}

```
      I       I       I       I
      v       v       v       v
// 1 2 3 // 4 5 6 |// 7 8 9 // 10 11 12 //
```

Se puede seguir un ejemplo utilizando la hoja de cálculo descrita en la figura I.

Este método tiene la ventaja de extender la muestra a toda la población. También recoge el posible efecto de estratificación. Por otra parte, este método tiene el inconveniente de que la población puede tener periodicidades que aumentan la homogeneidad interna de cada grupo. Lo que en el muestreo por conglomerados se denomina efecto de diseño. Y por tanto las estimaciones pueden tener un aumento del error debido al muestreo. Presenta también el inconveniente de que debido a esta periodicidad se plantea un problema para estimar el error debido al muestreo.

La semejanza con el muestreo estratificado viene dada de considerar a la población dividida en tantos estratos como unidades muestrales y seleccionar una unidad por estrato. En cuanto a la semejanza con el muestreo de conglomerados esta viene dada al considerar cada muestra sistemática como un conglomerado; y, por tanto, al muestreo sistemático como muestreo de conglomerados con una única unidad muestral.

Ejemplo de aplicación del método con intervalo de muestreo constante , utilizando una hoja de cálculo. En este ejemplo se forma el espacio muestral (I), se estiman las principales características de cada muestra (II) y sus errores de muestreo (III) sobre una población ficticia del tipo 1, 2,3, ...12. El tamaño de muestra es 4 y el tamaño del grupo para selección es 3.

Tabla 10.- Esquema sistemático con intervalo constante y orden aleatorio.

MUESTRAS	MEDIA[1]	VAR[2]	DES[3]	LI[4]	LS	CVS%	SIG2[5]	SIG	CUA[6]	CUD	CV	M2[7]	M3	M4
1 4 7 10	5.50	2.5	1.58	2.34	8.66	29%	11.25	3.35	15	3.87	70%	41.50	352.00	3164.50
2 5 8 11	6.50	2.5	1.58	3.34	9.66	24%	11.25	3.35	15	3.84	60%	53.50	494.00	4844.50
3 6 9 12	7.50	2.5	1.58	4.34	10.6	21%	11.25	3.35	15	3.84	52%	67.50	675.00	7168.50
MEDIA[8]	6.50	2.5	1.58	3.34	9.66	25%	11.25	3.35	15	3.84	61%	54.17	507.00	5059.17
VARIANZA	0.67	0.00	0.00	0.67	0.67	0%	0.00	0.00	0.00	0.00	1%	112.8	17472.6	2695043
VIEST/ DESVIACION	0.82	0.00	0.00	0.82	0.82	3%	0.00	0.00	0.00	0.00	8%	10.62	132.18	1641.66
COEF VAR	13%	0%	0%	24%	8%	13%	0%	0%	0%	0%	13%	20%	26%	32%
LIM SUP	8.13	2.5	1.58	4.97	11.3	31%	11.25	3.35	15	3.84	76%	75.42	771.37	8342.48
LIM INF	4.87	2.5	1.58	1.70	8.03	18%	11.25	3.35	15	3.84	45%	32.92	242.63	1775.85
POB (M.SR.)[9]	6.50	1.99	1.41	1.1	1.09	22%	11.92	3.45	17.8	4.23	65%	54.17	507.00	5059.17
POB (M.STR.)	6.50	0.16	0.4	5.79	7.39	6%								

[1] MEDIA es el estimador de la media calculado para cada muestra
[2] VAR representa el estimador de la varianza calculado para cada muestra
[3] DES error de muestreo. Raiz cuadrada de la expresión anterior.
[4] Li y Ls Limites de confianza al 95% supuesta normalidad.
[5] SIG2 Varianza poblacional. SiG es su raiz cuadrada.
[6] CUA Cuasivarianza poblacional. CUD raiz de CUA.
[7] M2-4 Momentos respecto del origen.
[8] Parámetros calculados sobre el conjunto de las muestras posibles.
[9] Parámetros calculados sobre el conjunto de población para el muestreo sin reposicion y después para el muestreo estratificado con una unidad por estrato.

En un círculo, ◯ , figuran los resultados que se comentan en el apartado correspondiente

El método de selección sistemática con intervalo de muestreo constante puede describirse de una manera formal utilizando la siguiente notación: La selección se lleva a cabo obteniendo un número aleatorio entre 1 y k, que denominaremos i. Este número permite determinar la primera unidad que figurará en la muestra. Las n-1 unidades restantes se seleccionan de forma que:

$zc = i + (j-1)\,k$

Donde

zc = Valor que identifica las unidades seleccionadas con intervalo constante

i = número aleatorio de selección. Entre 1 y k

j = número correlativo entre 1 y n. Donde n es el tamaño de muestra.

k = Tamaño de los grupos formados para la selección. $k = N/n$

5.2. MUESTREO SISTEMÁTICO CON INTERVALO VARIABLE

Supone una mejora del método con selección de intervalo constante. Se trata de un método que persigue aumentar la heterogeneidad interna de la muestra. Consiste en seleccionar la muestra con un intervalo variable que tenga como consecuencia un efecto de movilidad dentro de cada estrato.

La selección consiste en obtener aleatoriamente una primera unidad y seleccionar las siguientes añadiéndole un termino de movilidad. Para que todos los estratos estén representados, al factor de movilidad se le impone la restricción de que si la unidad seleccionada es igual al límite superior de un estrato automáticamente se le restan al contador k unidades, y comienza desde el borde siguiente el mismo proceso. Por ejemplo con una población de 12 unidades y un tamaño de muestra de 4 el proceso sería el siguiente:

N=4 ; n=3 ; k=4

Unidades de la Población

1 2 3 4

5 6 7 8

9 10 11 12

Si seleccionamos aleatoriamente la unidad 1, la siguiente será $U(1)+(k+1)=1+4+1=6$, Y la siguiente unidad seleccionada será $U(6)+(k+1)=11$, tendremos así seleccionada una muestra sistemática con intervalo variable y formada por las unidades (1,6,11). Si la primera unidad seleccionada hubiese sido, siguiendo un procedimiento aleatorio, la unidad 3, entonces la siguiente unidad seleccionada sería $U(3)+(k+1)=3+4+1=8$; para seleccionar la siguiente unidad el método exige, para que todas las unidades de la población tengan la misma probabilidad de ser seleccionadas, que el factor de movilidad se ponga a cero, y por tanto la unidad será $U(8)+(0+1)=9$. Así la muestra quedaría compuesta por las unidades (3,8,9).

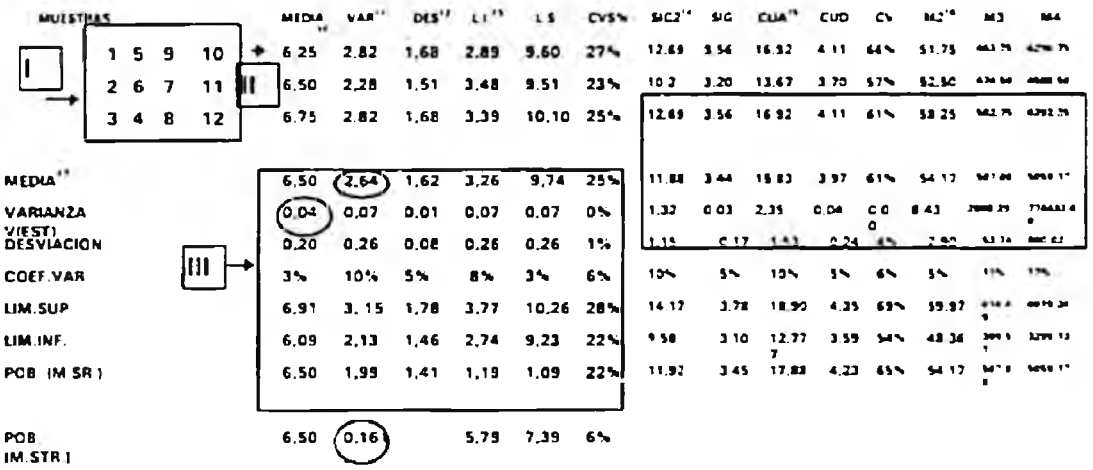
Vamos a obtener todas las muestras posibles, con tamaño de muestra igual a tres:

Utilizando intervalo constante				o intervalo variable			
	U(1)	U(2)	U(3)		U(1)	U(2)	U(3)
Muestra (1)	1	5	9	Muestra(1)	1	6	11
Muestra (2)	2	6	10	Muestra(2)	2	7	12
Muestra (3)	3	7	11	Muestra(3)	3	8	9
Muestra (4)	4	8	12	Muestra(4)	4	5	10

Los resultados de este método aplicado a los datos anteriores se pueden obtener utilizando el esquema de hoja de cálculo , mostrado en la figura 2.

Ejemplo de aplicación del método con intervalo de muestreo variable , utilizando una hoja de cálculo. En este ejemplo se forma el espacio muestral (I), se estiman las principales características de cada muestra (II) y sus errores de muestreo (III) sobre una población ficticia del tipo 1, 2,3, ...12. El tamaño de muestra es 4 y el número de grupos para selección 3.

Tabla 11.- Esquema sistemático con intervalo variable y orden aleatorio.



MUESTRAS	MEDIA[10]	VAR[11]	DES[12]	LI[13]	LS	CV%	SIG2[14]	SIG	CUA[15]	CUD	CV	M2-4[16]	M3	M4
1 5 9 10	6,25	2,82	1,68	2,89	9,60	27%	12,69	3,56	16,92	4,11	66%	51,75		
2 6 7 11	6,50	2,28	1,51	3,48	9,51	23%	10,2	3,20	13,67	3,70	57%	52,50		
3 4 8 12	6,75	2,82	1,68	3,39	10,10	25%	12,69	3,56	16,92	4,11	61%	58,25		
MEDIA[17]	6,50	2,64	1,62	3,26	9,74	25%	11,88	3,44	15,83	3,97	61%	54,17		
VARIANZA V(EST)	0,04	0,07	0,01	0,07	0,07	0%	1,32	0,03	2,35	0,04	0,00	8,43		
DESVIACION	0,20	0,26	0,08	0,26	0,26	1%	1,15	0,17	1,53	0,24	6%	2,90		
COEF.VAR	3%	10%	5%	8%	3%	6%	18%	5%	15%	5%	6%	5%		
LIM.SUP	6,91	3,15	1,78	3,77	10,26	28%	14,17	3,78	18,90	4,35	69%	59,97		
LIM.INF.	6,09	2,13	1,46	2,74	9,23	22%	9,58	3,10	12,77	3,59	54%	48,36		
POB (M.SR.)	6,50	1,99	1,41	1,19	1,09	22%	11,92	3,45	17,83	4,23	65%	54,17		
POB IM.STR.)	6,50	0,16		5,79	7,39	6%								

En un círculo, ◯ , figuran los resultados que se comentan en el apartado correspondiente

[10] MEDIA es el estimador de la media calculado para cada muestra
[11] VAR representa el estimador de la varianza calculado para cada muestra
[12] DES error de muestreo. Raíz cuadrada de la expresión anterior.
[13] Li y Ls Límites de confianza al 95% supuesta normalidad.
[14] SIG2 Varianza poblacional. SIG es su raiz cuadrada.
[15] CUA Cuasivarianza poblacional. CUD raíz de CUA.
[16] M2-4 Momentos respecto del origen.
[17] Parámetros calculados sobre el conjunto de las muestras posibles.

De una manera más formal y utilizando una notación similar a la del apartado anterior, tendríamos que:

$$zv = i + (j-1)(k+1) - c\,k$$

zv = Valor de la unidad muestral seleccionada con intervalo variable. En este caso coincide con el índice que identifica la unidad.

Donde, por definición, los valores de toma c vienen dados por las siguientes situaciones:

- Si zv = jk no ha sucedido nunca, entonces c = 0
- Si zv = jk ha ocurrido una vez, entonces c = 1
- Si zv = jk ha ocurrido dos veces, entonces c = 2 ....

Así, c = 0,1,2 ... de acuerdo con el número de veces que ha ocurrido que zv = jk

Se incluye un ejemplo con una población un poco mayor que la utilizada en la ilustración de la hoja de cálculo para comprender mejor el método propuesto:

A continuación tenemos la situación definida por un tamaño de población N = 36; un tamaño de muestra n = 9; un número aleatorio para la primera selección y = 2; tamaño de los grupos para la selección k = 4 y los valores muestrales seleccionados con este método serán:

Valores de la población	Número aleatorio i	Número correlativo j	Límite del grupo jk	Condición para el grupo c	Valores que forman la muestra i+(j-1)(k+1) -ck	
1 2 3 4	2	1	4	0 (*)	2 + (1-1)(4+1) - 0x4 =	2
5 6 7 8		2	8	0	2 + (2-1)(4+1) - 0x4 =	7
9 10 11 12		3	12	0	2 + (3-1)(4+1) - 0x4 =	12
13 14 15 16		4	16	1 (**)	2 + (4-1)(4+1) - 1x4 =	13
17 18 19 20		5	20	1	2 + (5-1)(4+1) - 1x4 =	18
21 22 23 24		6	24	1	2 + (6-1)(4+1) - 1x4 =	23
25 26 27 28		7	28	1	2 + (7-1)(4+1) - 1x4 =	28
29 30 31 32		8	32	2 (***)	2 + (8-1)(4+1) - 2x4 =	29
33 34 35 36		9	36	2	2 + (9-1)(4+1) - 2x4 =	34

(*) No ha ocurrido que zv=jv, así c=0.

(**) Ha ocurrido una vez que zv=jk, así c=1.

(***) Es la segunda vez que ocurre que zv=jk, así c=2.

Este método aplicado a datos del tipo $X_i = i$ para i=1,2,...N, con N cualquier número natural, presenta un caso de muestras perfectamente equilibradas cuando $n = \sqrt{N}$. En este caso el error debido al proceso de muestreo es cero. Esto es, sea cual sea la muestra seleccionada la estimación de la media coincide con la media poblacional.

5.3. PRÁCTICAS PARA MUESTREO SISTEMÁTICO VARIABLE

1.- Para una población de tamaño 100, donde los valores de la variable son $X_i = i$ con $i = 1, 2, 3, ..., 100$. Para muestras de tamaño 5, calcular el espacio muestral, los valores estimados para cada muestra utilizando la fórmula del error para el muestreo aleatorio sin reposición y los valores estimados en base a todas las muestras posibles.

2.- Con los datos del problema anterior aplicar el método sistemático con intervalo variable.

3.- Para la misma población comparar los datos del fichero histórico con los obtenidos al aplicar muestreo sistemático con intervalo variable.

5.4. PROBLEMA DE LA ESTIMACIÓN DEL ERROR DE MUESTREO

En la hoja de cálculo empleada en los ejemplos anteriores, podemos observar el problema que se presenta de estimación de varianzas al aplicar muestreo sistemático. Mientras que la varianza de la media es 0,04, la media de las varianzas es 2,64. Esto es debido a que las varianzas de cada muestra se obtienen considerando que el orden de la población es aleatorio, cuando menos cierto sea este supuesto más alejada estará la estimación de su verdadero valor. No obstante, se puede utilizar este método para estimar la varianza del estimador, si no se dispone de otro, en tanto que se trata de una estimación por exceso, proporciona una cota superior del verdadero error. En un capítulo posterior se darán varias alternativas para estimar la varianza en para el muestreo sistemático con intervalo variable: una da lugar a lo que se denominará muestreo sistemático con intervalo variable y estimación de la varianza con la técnica del lazo; otra alternativa consiste en calcular un factor de ajuste que utilizando la información obtenida por el proceso de reiteraciones, proporciona un estimador insesgado de la varianza; también se considerará la posibilidad de utilizar información proporcionada por el coeficiente de correlación intraconglomerados; por ultimo se puede considerar también la utilización de un estimador autogenerado.

El error debido al muestreo se puede calcular, si conocemos todas las muestras posibles, con la fórmula de la desviación típica donde los valores son las estimaciones y la media su valor esperado o valor poblacional. Ahora bien, si solamente conocemos una muestra la estimación del error se puede llevar a cabo aplicando la fórmula del muestreo aleatorio simple que proporcionará una estimación por exceso, puesto que es previsible que al aplicar el método del muestreo sistemático el error sea inferior, siempre que no existan correlaciones no deseadas en la población.

5.4.1. SUPUESTO QUE EL ORDEN DE LA POBLACIÓN ES ALEATORIO

Considerando una población de 400 unidades y un tamaño de muestra igual a 16, el error de muestreo, calculado con la información de cada muestra, proporciona ,en el caso del muestreo sistemático con intervalo constante, una estimación invariable de 29.1. El muestreo sistemático con intervalo variable proporciona una estimación variable en torno al 29,2 con una desviación de 1.07. Esto es debido a la periodicidad constante en un caso y móvil en el segundo caso. El resultado es una estimación en media inferior para cada muestra cuando el muestreo es con intervalo constante que cuando no lo es. Este resultado es debido a que con el procedimiento de selección sistemática con intervalo variable se ha aumentado la heterogeneidad interna de la muestra, o lo que es lo mismo al ordenar los valores de la población dos unidades consecutivas en la muestra son más heterogéneas entre sí cuando se utiliza intervalo variable que cuando se utiliza intervalo constante.

Al aumentar el tamaño de muestra de 16 a 20 el resultado se puede ver en el gráfico de la aplicación. En este caso los dos métodos proporcionan una estimación similar, sin embargo el error de muestreo real, esto es el calculado con todas las muestras posibles, es cero para el caso de solución de intervalo variable. Este aspecto se puede comprender mejor en el apartado de tamaño óptimo de muestra o simplemente observando el gráfico con las estimaciones para este caso.

Resultados similares se pueden obtener para poblaciones de 800 unidades y muestras de 16 o 32 elementos. En cualquier caso la estimación obtenida de la muestra deberá considerarse como una estimación por exceso. Y al evaluar el método será mejor utilizar la estimación del error real, calculada utilizando la información de todas las muestras posibles.

5.4.2. SUPUESTO QUE SE DISPONE DE INFORMACIÓN POBLACIONAL

En la aplicación POSDEM se ha implementado un procedimiento para recalcular el error debido al muestreo en cada muestra teniendo en cuenta la información suministrada por el proceso de reiteraciones. El procedimiento consiste en ponderar los errores debidos al muestreo calculados en cada muestra por un factor de ajuste obtenido de la siguiente expresión:

$$FA = \frac{VR}{EVM} = \frac{\sum_{i=1}^{k} (\bar{\bar{x}}_i - \bar{\bar{x}})^2}{\sum_{i=1}^{k} V(\bar{\bar{x}}_i)}$$

- VR = La varianza basada en el cálculo completo del espacio muestral o, en su defecto, en una representación del mismo utilizando reiteraciones.
- EVM = El valor esperado sobre todas las reiteraciones del estimador para la varianza en cada muestra.
- k = Reiteraciones.

Esta alternativa permite disponer de un estimador insesgado del error de muestreo en cada muestra. Tiene el inconveniente de precisar información censal que permita el proceso de reiteraciones para cada variable para la que se quiere calcular el error de muestreo.

5.4.3. UTILIZACIÓN DE UN ESTIMADOR AUTOGENERADO

Una posible vía de investigación, utilizando la aplicación POSDEM, consiste en aplicar esta alternativa sobre una muestra, como si se tratase de información censal. Para esto, para que una de las muestras seleccionadas se incorpore a un fichero que permita a su vez ser tratado como un listado censal se ha habilitado una opción en el menú de listados que se ha denominado "Obtener muestra", que permite exportar los datos de una determinada selección a un fichero externo a la aplicación.

De esta forma con el menú de cálculos con las alternativas "Bootstrap" y "Jackknife" se pueden establecer estimaciones del error debido al muestreo con estos procedimientos.

Se ha considerado también otra opción, que, con una filosofía parecida a la anterior, esto es considerando una muestra como si fuese una población, generando submuestras de tamaño dos.

5.4.3.1. PLAN DE TUKEY

Para exponer este apartado seguiremos al profesor Azorin: una solución, debida a Tukey [véase Deming (1950), y Jones (1956)) consiste en obtener no una sola muestra sistemática, sino varias que resultan al tomar aleatoriamente y sin reemplazamiento varios orígenes entre *l* y *k*, ambos inclusive.

En realidad pueden considerarse dos procedimientos. En el primero, que Jones denomina método *A*, se empieza calculando el intervalo de muestreo que para *m* muestras será $k=(mN)/n$, siendo *n* el tamaño general de la muestra.

Dividida la población en partes, la primera de las cuales comprende sus *k* primeros elementos y sucesivamente la segunda los *k* siguientes, etc., se seleccionan de cada parte sin reemplazamiento *m* elementos. Frecuentemente se toma $m=10$.

En la primera parte, los *n* elementos se seleccionan con una tabla de números aleatorios. En las otras, se suman múltiplos de *k* a los números de orden de los elementos elegidos en la primera parte. Si estos fueran *a, b, c, ..., j*, los de la segunda parte serán $a + k, b + k, \ldots, j + k$; los de la tercera $a + 2k, b + 2k, \ldots, j + 2k$, y así sucesivamente. La primera submuestra constará de los elementos, $a, a + k, a + 2k$, la segunda de $b, b + k, b + 2k$ ... y la *m*-ésima de los $j, j + k, j + 2k$.

En cuanto al método *B*, consiste en elegir aleatoriamente, también sin reemplazamiento, los números de cada parte, sin añadir múltiplos de *k* a los seleccionados en la primera. Solo difiere, pues, del anterior en el número de selecciones aleatorias. Las *m* submuestras se obtienen asignando números aleatorios a los elementos de la población, en el mismo orden en que han sido obtenidos en la tabla.

En los dos procedimientos anteriores se estima la media poblacional calculando la media aritmética de las *m* medias de la submuestra. Esto es, llamando *i1, i2, ..., im* a los *m* orígenes y representado por $\bar{Y}_{ij}$ la media correspondiente al *j*-ésimo origen, tendramos:

$$\bar{y}_i = \frac{\bar{y}_{i1}+\ldots+\bar{y}_{im}}{m}$$

La varianza será ahora:

$$\sigma^2_{\bar{y}} = \frac{k-m}{km}\,\frac{1}{k-1}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2$$

[Véase, p. ej., Gautschi (1957)], y como estimador de esta varianza puede tomarse:

$$\hat{\sigma}^2_{\bar{y}} = \frac{N-n}{N}\,\frac{1}{m(m-1)}\sum_{h=1}^{m}(\bar{y}_{ih} - \bar{y}_i)^2$$

[véase Jones (1956)]

5.4.4. UTILIZACIÓN DE DOS UNIDADES POR ESTRATO

Es posible, para estimar la varianza del estimador en diseños sistemáticos, utilizar el criterio de considerar dos unidades muestrales contiguas como parte de una muestra estratificada de tamaño dos. Este método se utilizará también para estimar la varianza cuando el mismo se realice conforme al procedimiento de muestreo aleatorio estratificado con una unidad muestral por estrato. Este es el procedimiento que la aplicación utiliza por defecto cuando el método de selección es sistemático.

5.5. TAMAÑO DE MUESTRA ÓPTIMO PARA EL MUESTREO SISTEMÁTICO CON INTERVALO VARIABLE

Existen dos formas básicas de optimizar una investigación por muestreo: disminuir el error de muestreo hasta un nivel que se considera aceptable y fijar el tamaño de muestra que se asocia con dicho error; o bien disminuir el tamaño de muestra hasta un nivel de coste que se considera aceptable y calcular el error de muestreo de dicho tamaño.

Cuando se aplica un método de muestreo sobre el que existe una teoría apropiada es posible, dado un error de muestreo y una variabilidad de los datos poblacionales, determinar con exactitud el tamaño de muestra necesario. En el caso de muestreo sistemático y por las dificultades que presenta para estimar el error de muestreo esto no es posible al menos con la exactitud deseada.

Para poblaciones del tipo $X_i = i$ para $i = 1,2,3..N$, en el caso de la solución de intervalo variable, el tamaño de muestra óptimo fijado el error de muestreo y sin hacer una consideración de los costes, coincide con la raíz cuadrada del tamaño de población. Teniendo en cuenta que si la cifra resultante no es múltiplo de la población utilizar una selección circular. En el caso de que el tamaño de muestra sea el primer múltiplo raíz cuadrada de la población entonces el error debido al muestro es

cero. Esto quiere decir que, independientemente de cual sea la muestra elegida con este procedimiento, la estimación del parámetro es siempre la misma y coincide con el valor del parámetro poblacional.

5.5.1. PRÁCTICAS PARA EL TAMAÑO DE MUESTRA ÓPTIMO Y POBLACIONES $X_i = i$ PARA $i = 1,2,...N$

Con el programa POSDEM se puede seguir la explicación del ejemplo siguiente: se ha considerado una población de 200 elementos y un tamaño de muestra de 20 unidades. El número total de muestras posibles es de 10, que coincide con el número de unidades en cada estrato. En el gráfico se puede observar como las estimaciones que proporciona el método de muestreo sistemático tradicional con intervalo constante oscilan entre 95 y 105. Así la muestra número 4 proporciona una estimación de 99 y la muestra número 9 una estimación de 104. Las muestras mejores son las que utilizan valores centrales, en este caso las muestras 5 y 6. El mismo método de selección, pero con la variación de realizar un intervalo variable proporciona, en este caso, el mismo valor para el estimador de la media independientemente de las muestra seleccionada. Cualquiera que sea la muestra obtenida proporciona la misma estimación, que coincide con el valor del parámetro que se desea estimar.

Las cuestiones planteadas en este apartado sobre tamaño de muestra óptimo, dado que en la práctica no se dan poblaciones del tipo 1,2, .. N, tiene exclusivamente un carácter de primera aproximación. Esto es, cuando nos encontramos con una población marco de 784 viviendas donde el valor de la variable es el número de personas por hogar, al aplicar muestreo sistemático con intervalo variable no podremos definir a priori un tamaño de muestra asociado a un determinado error de

muestreo, al no disponer de una expresión que relacione ambas expresiones, tamaño de muestra y error. Así, y con el único propósito de una primera aproximación obtendremos un tamaño de muestra igual a raíz cuadrada de 784, esto es el tamaño de muestra será de 28 viviendas. Una vez seleccionada la muestra podremos con la aplicación POSDEM, calcular el error de muestreo y decidir si éste se debe aumentar o disminuir.

5.6. ANÁLISIS DE LOS RESULTADOS COMPARACIÓN DE LOS MÉTODOS PROPUESTOS CON EL MUESTREO ESTRATIFICADO DE TAMAÑO DE MUESTRA UNO POR ESTRATO

A efectos de la simulación que se realiza para comprobar la bondad del método de intervalo variable se va a realizar un análisis de los errores de muestreo de los métodos propuestos. El marco utilizado ha sido la población descrita en el ejemplo resuelto con la hoja de cálculo. En esta presentación se han ampliado algunos cálculos al utilizar las posibilidades que permite la aplicación POSDEM.

Para intervalo de muestreo constante

$\rho = -0.25$

$S_h^2 = 4$

$S_w^2 = 15$

$V(\text{imc})_{sis} = 0.66$

$$C_v = \frac{\sqrt{V(imc)_{sis}}}{\bar{X}} \times 100 = 12.5\ \%$$

Para intervalo de muestreo variable

$\rho = -0.32$

$S_h^2 = 0.25$

$S_w^2 = 15.83$

$V(imv)_{sis} = 0.04$

$$C_v = \frac{\sqrt{V(imv)_{sis}}}{\bar{X}} \times 100 = 3.1\ \%$$

Al efecto de observar en una única tabla la ganancia o perdida de precisión que se produce al utilizar un método de selección de muestras en lugar de otro, se puede plantear el siguiente esquema: hemos visto que los errores de muestreo en términos de varianza para los cuatro métodos considerados han sido

Cv (Msr) = 22% Cv (Msis IMV) = 3%

Cv (Msis IMC) = 13% Cv (Mest nh = 1) = 6%

No se detallan las abreviaturas por ser evidentes.

De aquí se puede deducir la siguiente tabla:

Tabla 12.- Comparación de ganancias en precisión.

A usar / De utilizar	M_sr	M_sis_imc	M_sis_imc	M_est_nh = 1
M_sr	0%	-69%	-633%	-267%
M_sis_imc	41%	0%	-333%	-117%
M_sis_imc	86%	77%	0%	50%
M_est_nh = 1	73%	54/	-100%	0%

En él se define la ganancia (valores positivos) o perdida (valores negativos) de precisión del método (1) sobre el método (2) mediante la expresión:

$$G_{1\_2} = (1 - \frac{C_{v1}}{C_{v2}})100$$

La interpretación de estos valores pasa por considerar que, de utilizar un método a utilizar otro, se produce una ganancia (+) o perdida (-) de precisión en tanto por ciento del valor en precisión del método comparado para sustituir. Así , por ejemplo, de utilizar el método de muestreo sistemático con intervalo de selección variable

que tiene una precisión del 3% a utilizar muestreo estratificado con una unidad muestral por estrato, que tiene una precisión del 6 %, se produce una ganancia en precisión del 50 % respecto del mismo 6%.

El caso más interesante será aquel en el que todos los valores de la fila correspondientes a un determinado método sean todos positivos. Querrá decir que el método presenta una ganancia en precisión sobre cualquiera de los otros métodos considerados. En este ejemplo se verifica que el método que mejora en precisión a los restantes es el de muestreo sistemático con intervalo variable.

Para la población considerada se puede observar como la ganancia en precisión al utilizar muestreo sistemático con intervalo variable frente al intervalo constante reduce el error relativo de 12.5% a 3.1% , lo que parece un resultado prometedor. Es también otro resultado interesante el que el método con intervalo variable es más preciso, incluso, que el método estratificado con una unidad por estrato, que pasa de tener un error relativo del 6.1% al referido 3.1%.

Uno de los motivos para desarrollar la aplicación informática POSDEM es superar la limitación que supone el trabajar con poblaciones $X_i = i$ para $i = 1,2,3...N$. Con el programa para ordenador, POSDEM, es posible aplicar estos métodos y otros, y evaluarlo para el caso de poblaciones reales, tanto para variables cuantitativas como para variables cualitativas. Se han obtenido, como era de esperar, buenos resultados de una ilustración con datos reales sobre la Encuesta de Población Activa y la Encuesta Industrial.

CAPÍTULO 6

MUESTREO SISTEMÁTICO II

6.1. ESTIMACIÓN DEL ERROR CON LA TÉCNICA DEL LAZO

Si suponemos conocida, por una operación censal anterior o por la existencia de un registro administrativo, una variable de interés de la población, en este caso para la población completa, esta información puede incorporase al proceso de selección de una muestra aleatoria de distintas formas. Habitualmente se agrupan las unidades de la población de forma que los grupos sean lo mas homogéneos posibles en su interior. De esta forma con pocas unidades de cada grupo se podrá tener una estimación precisa de los valores que caracterizan a la población. Esta forma de agrupar unidades se denomina diseño muestral estratificado. Aquí, no obstante se va utilizar la expresión estratificación para denominar el procedimiento que consiste en unir, una vez ordenadas las unidades de la población con esa variable censal, dos unidades muestrales próximas y disponer del máximo número de estratos posible con un tamaño de muestra por estrato igual a dos. De esta forma se puede estimar el error del muestreo sistemático mediante las ponderaciones del muestreo estratificado, con los datos de una sola muestra.

6.1.1. APLICADO AL MUESTREO SISTEMÁTICO CON INTERVALO CONSTANTE

Para una población de tamaño 12, tamaño de muestra 4 y número de reiteraciones 3, se pide: obtener las muestras posibles, sus estimadores y calcular el error de muestreo aplicando las ponderaciones del muestreo estratificado.

Vamos a considerar dos casos: en el primero la muestra sistemática se ha obtenido con intervalo constante y en el segundo con intervalo variable. Las dos hojas de cálculo siguientes muestran los resultados de cada problema. También se puede seguir el ejemplo con la aplicación POSDEM.

Tabla 13.- Esquema sistemático con intervalo constante y técnica del lazo.

		MEDIA[1]	VAR[2]	DES[3]	LI[4]	LS	CVS%	SIG2[5]	SIG	CUA[6]	CUD	CV	M2[7]	M3	M4
	1 4 7 10	5,50	0,75	0,87	3,	7,	16%	11,25	3,35	22,50	4,74	86%	41,50	352,00	3164,50
	2 5 8 11	6,50	0,75	0,87	4,	8,	13%	11,25	3,35	22,50	4,74	73%	53,50	494,00	4844,50
	3 6 9 12	7,50	0,75	0,87	5,	9,	12%	11,25	3,35	22,50	4,74	63%	67,50	675,00	7168,50
MEDIA[8]		6,50	0,75	0,87	4,77	8,23	14%	11,25	3,35	22,50	4,74	0,74	54,17	507,00	5058,17
VARIANZA		0,67	0,00	0,00	0,67	0,67	0%	0,00	0,00	0,00	0,00	0,00	112,3	17472,6	2695043
DESVIACION		0,82	0,00	0,00	0,82	0,82	2%	0,00	0,00	0,00	0,00	9%	10,62	132,18	1641,66
COEF.VAR		13%	0%	0%	17%	10%	13%	0%	0%	0%	0%	13%	20%	26%	32%
LIM.SUP		8,13	0,75	0,87	6,40	9,87	17%	11,25	3,35	22,50	4,74	93%	75,42	771,37	8342,48
LIM.INF.		4,87	0,75	0,87	3,13	6,60	10%	11,25	3,35	22,50	4,74	55%	32,92	242,63	1775,85

En este caso el error asociado con cada muestra es siempre constante e igual a 0,75; se puede observar que se aproxima bastante al verdadero valor calculado en base a las reiteraciones 0,67

Hay que recordar que cuando se utilizaba para estimar el error de muestreo el supuesto de que el orden de la población es aleatorio la media de las varianzas del estimador calculadas en cada muestra eran 2.5 para intervalo constante y 2.64 para intervalo variable.

---

[1] MEDIA es el estimador de la media calculado para cada muestra
[2] VAR representa el estimador de la varianza calculado para cada muestra
[3] DES error de muestreo. Raíz cuadrada de la expresión anterior.
[4] Li y Ls Límites de confianza al 95% supuesta normalidad.
[5] SIG2 Varianza poblacional. SiG es su raiz cuadrada.
[6] CUA Cuasivarianza poblacional. CUD raiz de CUA.
[7] M2-4 Momentos respecto del origen.
[8] Parámetros calculados sobre el conjunto de las muestras posibles.

6.1.2. APLICADO AL MUESTREO SISTEMÁTICO CON INTERVALO VARIABLE

Los resultados para el caso de muestreo sistemático con intervalo variable

Tabla 14.- Esquema sistemático con intervalo variable y técnica del lazo.

MUESTRAS	MEDIA[9]	VAR[10]	DES[11]	LI[12]	LS	CVS %	SIG2[13]	SIG	CUA[14]	CUD	CV	M2[15]	M3	M4
1 5 9 10	6,25	0,71	0,84	4,	7,	13%	12,69	3,56	25,38	5,04	81%	51,7	463,	4296,7
2 6 7 11	6,50	1,33	1,15	4,	8,	18%	10,25	3,20	20,50	4,53	70%	52,5	474,	4588,5
3 4 8 12	6,75	0,71	0,84	5,	8,	12%	12,69	3,56	25,38	5,04	75%	56,2	582,	6292,2
MEDIA[16]	6,50	0,92	0,95	4,61	8,39	15%	11,88	3,44	23,75	4,87	0,75	54,1	507,0	5059,1
VARIANZA	0,04	0,09	0,02	0,13	0,13	0%	1,32	0,03	5,28	0,06	0,00	8,43	2888	774433
DESVIACION	0,20	0,29	0,15	0,36	0,36	2%	1,15	0,17	2,30	0,24	4%	2,90	53,7	880,02
COEF VAR	3%	32%	16%	8%	4%	16%	10%	5%	10%	5%	6%	5%	11%	17%
LIM SUP	6,91	1,51	1,24	5,33	9,11	19%	14,17	3,78	28,35	5,35	84%	59,9	614,	6819,2
LIM INF	6,09	0,33	0,65	3,89	7,67	10%	8,58	3,10	19,15	4,39	66%	48,3	399,	3299,1

El error asociado con cada muestra se mueve en torno a 0,92. Esto es, ha aumentado respecto del método anterior. Sin embargo y lo que es más importante el error basado en las reiteraciones ha

[9] MEDIA es el estimador de la media calculado para cada muestra
[10] VAR representa el estimador de la varianza calculado para cada muestra
[11] DES error de muestreo. Raíz cuadrada de la expresión anterior.
[12] Li y Ls Límites de confianza al 95% supuesta normalidad.
[13] SIG2 Varianza poblacional. SIG es su raíz cuadrada.
[14] CUA Cuasivarianza poblacional. CUD raíz de CUA.
[15] M2-4 Momentos respecto del origen.
[16] Parámetros calculados sobre el conjunto de las muestras posibles.

disminuido de 0,67 a 0,04. Esto es debido a que con el método variable aumenta la heterogeneidad interna de la muestra (pasa de 0.75 a 0.92) y por tanto también aumenta la homogeneidad entre las muestras (por eso pasa de 0.67 a 0.04). Por tanto, con el método de intervalo variable sea cual sea la muestra los resultados de la estimación variarán poco en relación con su verdadero valor poblacional, que es 6,5. El inconveniente se encuentra en no disponer de un método para estimar el error en base a los datos de una sola muestra. En este caso como en el anterior, cuando se supone que el orden de la población es aleatorio, el utilizar las ponderaciones del muestreo estratificado con dos unidades por estrato pueden considerarse como una estimación todavía por exceso.

6.1.3. APLICADO AL MUESTREO SISTEMÁTICO EQUILIBRADO Y MODIFICADO

- Si en el caso de muestreo de dos conglomerados formados sistemáticamente con intervalo constante los principales resultados, para la población ficticia que se ha considerado eran:

Varianza del estimador basada en el conjunto de las muestras posibles = 0,67

Varianza del estimador basada en cada muestra. Orden aleatorio. Valor esperado = 2,5

Varianza del estimador basada en cada muestra. Técnica del lazo. Valor esperado = 0,75

- Para el caso de muestreo de dos conglomerados formados sistemáticamente con intervalo constante son:

Varianza del estimador basada en el conjunto de las muestras posibles = 0,04

Varianza del estimador basada en cada muestra. Orden aleatorio. Valor esperado = 2,64

Varianza del estimador basada en cada muestra. Técnica del lazo. Valor esperado = 0,92

- Para el caso de muestreo de dos conglomerados equilibrados :

Varianza del estimador basada en el conjunto de las muestras posibles = 0,00

Varianza del estimador basada en cada muestra. Orden aleatorio. Valor esperado = 2,65

Varianza del estimador basada en cada muestra. Técnica del lazo. Valor esperado = 0,97

- Para el caso de muestreo de dos conglomerados modificados:

Varianza del estimador basada en el conjunto de las muestras posibles = 0,00

Varianza del estimador basada en cada muestra. Orden aleatorio. Valor esperado = 2,65

Varianza del estimador basada en cada muestra. Técnica del lazo. Valor esperado = 3,97

De esta tabla de valores puede observarse como varía la dificultad para estimar el error debido al muestreo para los métodos y datos considerados, al emplear la técnica del lazo para estimar el error con los datos de una sola muestra.

6.1.4. PRÁCTICAS

Comparar los métodos de intervalo constante y variable para los siguientes valores:

1.-) Población N=400 , Muestra n=20 y reiteraciones RT=20.

2.-) Población N=400 , Muestra n=10 y reiteraciones RT=40.

6.2. EJEMPLOS CON DATOS CUALITATIVOS

En la aplicación POSDEM se ha implementado la posibilidad de procesar datos cualitativos del tipo 0,1. Así, si suponemos una población de 16 unidades, identificadas por un número correlativo, y que en cada una de ellas se ha investigado la ausencia o presencia de un cierto carácter; por tanto, en este caso el valor de la variable sólo podrá tomar valores cero y uno, así tendremos:

Ui =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Vv(Ui) =	1	0	1	0	1	0	0	1	0	1	1	1	1	0	0	1

Este fichero se ha denominado edc1.dbf y está disponible en la aplicación POSDEM para realizar ejemplos con él. En este caso si aplicamos un muestreo sistemático con intervalo variable, fijando un tamaño de muestra igual 4, tendremos un espacio muestral que proporcionará para cada muestra las siguientes estimaciones de la varianza para el estimador de la proporción y estimación del error con la "técnica del lazo".

A efectos de comprobación de los cálculos realizados se puede seguir la siguiente tabla:

Espacio muestral $S(\underline{x})$	Proporción $\hat{P}$	Varianzas para el estimador de la proporción $\hat{V}(\hat{p}_{st}) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 (1-\frac{n_h}{N_h})(\hat{p}_h(1-\hat{p}_h))$
		(8/16)^2 x (1-2/8) x .5 x .5 + .18 x 1 x .0
1 0 1 1	.75	.046
0 0 1 1	.5	.0
1 1 0 0	.5	.0
0 1 1 0	.5	.0936

Otro ejemplo, para una población de 24 unidades y tamaño de muestra igual a 6 toma los valores poblacionales, incluidos en el fichero edc2.asc, siguientes:

Ui = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Vv(Ui)= 1 0 1 0 1 0 0 1 0 1 1 1 1 0 0 1 0 1 0 1 1 1 0 0

Así, el espacio muestral y las estimaciones de la varianza para el estimador de la proporción serán en este caso:

A efectos de comprobación de los cálculos realizados se puede seguir la siguiente tabla:

Espacio muestral $S(\underline{x})$	Proporción $\hat{P}$	Varianzas para el estimador de la proporción $\hat{V}(\hat{p}_{st}) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 (1-\frac{n_h}{N_h})(\hat{p}_h(1-\hat{p}_h))$
1 0 1 1 0 1	.66	.046 + .18 x .5 x .5 = .091
0 0 1 1 1 0	.5	.0 + .18 x .5 x .5 = .045
1 1 0 0 0 0	.33	.0 + .18 x .0 = .0
0 1 1 0 1 1	.66	.0938 + .18 x 1 x 1 x .0 = .0938

Estos ejemplos pueden seguirse de forma más completa mediante la aplicación POSDEM, utilizando los datos externos a la aplicación contenidos en los ficheros edc1.dbf y edc2.dbf.

6.3. MUESTREO DE DOS CONGLOMERADOS FORMADOS SISTEMÁTICAMENTE

Otra variante, inversa a la estratificación, es formar en la población los grupos de unidades de manera que sean lo más heterogéneos entre sí. De esta forma seleccionando un solo grupo completo se dispone de una muestra que puede proporcionar estimaciones precisas. Este método se denomina muestreo de conglomerados. En la aplicación POSDEM este método se define con dos conglomerados por muestra, y permite obtener estimaciones y su error de muestreo. Los conglomerados se obtienen con el procedimiento de selección sistemática, por eso se ha denominado muestreo de dos conglomerados formados sistemáticamente. Por supuesto este método es distinto al de muestreo sistemático de conglomerados.

Para comprender mejor cómo se realiza en este caso el agrupamiento de unidades, vamos a seguir el siguiente ejemplo; datos: Población de tamaño 16, muestra igual a 4 y reiteraciones igual a 4. Si queremos tener un tamaño de muestra final igual a 4, y dado que en la aplicación POSDEM el número de conglomerados, por definición del método se ha fijado para optimizar el procedimiento siempre en un valor igual a 2, por ello tendremos que fijar un tamaño de muestra inicial, en este caso, también igual a 2. Desde el punto de vista de la computación este es un proceso invisible al usuario. El programa se ocupa entonces de

obtener las muestras posibles de tamaño dos con intervalo constante o variable y de casar posteriormente la primera muestra con la que ocupa la reiteración RT/2 y sucesivamente, hasta tener configurado un espacio muestral de reiteraciones igual a RT/2 y tamaño de muestra final igual a (tamaño de muestra inicial) x 2 = 4 (tamaño de muestra final).

Posteriormente para calcular las estimaciones del error de muestreo se utilizan las ponderaciones del muestreo estratificado, o bien mediante una de las opciones del programa con un procedimiento que tiene en cuenta la definición de varianza entre conglomerados Se reproducen las dos pantallas más significativas con los listados obtenidos para cada muestra y para el conjunto de las muestras posibles.

POSDEM

Ficha Técnica de la Encuesta: datos del diseño

Población marco utilizada: X, =i para i=1,2,...16
variable de estudio: valores del 1 al 16, coinciden con la identificación de la unidad poblacional.
Variable auxiliar: se ha ordenado la población por valor de la variable de estudio.

Método de muestreo empleado : de dos conglomerados formados sistemáticamente con intervalo variable

Tamaño de población = 16
Tamaño de muestra = 4
Número de muestras posibles = 4

Media poblacional = 8,5
Desviación típica (n-1) = 22,67

Fracción de muestreo = 25%
Desviación típica del estimador = ,71
Error de muestreo CV% (Basado en todas las muestras)= 8,32%

POSDEM

_ muestreo de dos conglomerados formados sistemáticamente con intervalo variable

_Listado de unidades seleccionadas en cada muestra
_Valor de la variable de identificación

Notas:

1.- En filas se representa cada muestra obtenida, en columnas los valores muestrales.

2.- Ejemplo: la posición Muestra(4),u(2) representa el valor correspondiente a la unidad
2 de la muestra 4.

3.- Muestras obtenidas = 4 Tamaño de muestra = 4 Tamaño de población = 16

	u(1)	u(2)	u(3)	u(4)
Muestra(1)	1	10	5	14
Muestra(2)	2	11	6	15
Muestra(3)	3	12	7	16
Muestra(4)	4	13	8	9

POSDEM

_ muestreo de dos conglomerados formados sistemáticamente con intervalo variable

_Estimadores para cada muestra

Notas:

1.- En filas se representa, la media, varianza, ... calculadas para todas las muestras obtenidas. En columnas figura el estimador al que se refieren los cálculos. Ejemplo, el valor media, varmed representa la media de todas las varianzas del estimador media, calculada en base a todas las muestras.

2.- Estimadores sesgados de S y CVS, corregidos. En la línea correspondiente figuran los
valores de S y de CVS calculados como raíz de S^2. El resto de los cálculos, por ejemplo
varianza de S se han mantenido en su formato original

	MEDIA	VARMED	DESMED	LIMED	LSMED	CVSMED(%)	
MEDIA	8.50	6.66	2.58	3.34	13.66	30.35	
VARIANZA...	.50	2.64	.12	.97	.97	25.00	
DESVIACION	.71	1.62	.34	.99	.99	5.00	
CO.VAR .(%)	8.32	24.40	13.47	29.15	7.25	16.50	
LIM.SU....		9.91	9.90	3.25	5.36	15.59	40.31
LIM.IF....	7.09	3.41	1.87	1.41	11.64	20.31	

POSDEM 17/06/97 19:53:26

_ muestreo de dos conglomerados formados sistemáticamente con intervalo variable

_Estimadores para cada muestra

Notas:

1.- En filas se representa cada muestra obtenida, en columnas los valores estimados.

2.- Ejemplo: el valor muestra(4),varmed representa la varianza del estimador media calculado para la muestra 4.

3.- Muestras obtenidas = 4 Tamaño de muestra = 4 Tamaño de población = 16

Sugerencia:

1.- Compruebe como los estimadores de cada muestra oscilan entre los intervalos de confianza calculados para el conjunto de todas las muestras. Para esto seleccione primero la opción cálculos para todas las muestras, posteriormente seleccione la opción cálculos para cada muestra.

	MEDIA	VARMED	DESMED	LIMED	LSMED	CVSMED(%)
Muestra(1)	7,5	7,59	2,75	1,98	13,01	36,74
Muestra(2)	8,5	7,59	2,75	2,98	14,01	32,41
Muestra(3)	9,5	7,59	2,75	3,98	15,01	29,00
Muestra(4)	8,5	3,84	1,96	4,57	12,42	23,06

En este último listado se puede comprobar la dificultad de estimar el error de muestreo con los datos de una sola muestra.

En definitiva en este apartado se proponen cuatro nuevos métodos de selección de muestras que son variantes de los métodos sistemático con intervalo constante, variable, sistemático equilibrado y modificado. Su efectividad depende al igual que en los restantes casos de la estructura poblacional que se este utilizando como marco.

6.4. PRÁCTICAS PARA EL MUESTREO DOS CONGLOMERADOS

1.- Obtener con los datos del ejemplo anterior los resultados para el caso de muestreo de dos conglomerados formados sistemáticamente con intervalo constante. Comentar los resultados comparándolos con el caso de muestreo de dos conglomerados formados sistemáticamente con intervalo variable.

2.- Obtener un gráfico del fichero histórico donde se comparen con esos mismos datos los resultados correspondientes a los ocho principales métodos sistemáticos utilizados en la aplicación POSDEM. Utilizar una población de tamaño 40 y muestra 8.

CAPÍTULO 7

COEFICIENTE DE CORRELACIÓN INTRACLÁSICA

7.1. ESTUDIO DE LOS VALORES DEL COEFICIENTE DE CORRELACIÓN INTRACLÁSICA Y DE LA VARIANZA DEL ESTIMADOR EN DISEÑOS SISTEMÁTICOS

Puede encontrase una introducción muy aclaratoria al concepto que mide el coeficiente de correlación intraclásica en Kendall,1947:303-308. Su expresión matemática viene dada por :

$$\rho(k = N \mid n.x) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} \sum_{t=j+n}^{n} ((x_{i,j} - \bar{X})(x_{i,t} - \bar{X}))}{\sigma^2} * \frac{2}{k((n(n-1))}$$

El coeficiente de correlación intraclásica también se puede presentar como coeficiente de correlación serial en la forma que puede verse en Cansado, 1950:96-97.

En esta sección obtendremos los límites para una aproximación de la varianza del estimador en el muestreo sistemático. Esta aproximación se llevará a cabo en función del valor del coeficiente de correlación intraclásico, el tamaño de la muestra y la varianza del muestreo con o sin reposición.

a.- Es claro que si el coeficiente de correlación intraclásica es igual a la expresión (-1/ n), entonces la varianza en muestreo

sistemático es igual a la varianza en el muestreo aleatorio con reposición dividido por el tamaño de muestra.

En el muestreo sistemático es un resultado conocido:

$$V(\bar{x}_s) = \frac{\sigma^2}{n}[1 + (n-1)\rho]$$

y que para el muestreo aleatorio con reposición es

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

De ambas expresiones tenemos:

$$V(\bar{x}_s) = V(\bar{x})[1 + (n-1)\rho$$

Si asumimos que

$\frac{\hat{V}(\bar{x})}{n} \equiv V(\bar{x}_s)$ es una buena aproximación.

Tendríamos

$$V(\bar{x})[1 + (n-1)\rho] = \frac{V(\bar{x})}{n}$$

y de aquí $\rho = \left(\frac{1}{n} - 1\right)\frac{1}{(n-1)} = \frac{-1}{n}$

Así, cuando

$$\rho = \frac{-1}{n}$$

podríamos usar la expresión $\frac{\hat{V}(\bar{x})}{n}$ como un indicador de $V(\bar{x}_s)$.

Este valor de ρ se representará por ρ_1.

b.- El valor del coeficiente de correlación intraclásica que hace cero la varianza en el muestreo sistemático es:

$\rho = \left(\frac{-1}{n-1}\right)$ ya que entonces $V(\bar{x}_s) = 0$

Este valor de ρ se representará por ρ_0.

c.- Si $\rho > 0$ la varianza en el muestreo aleatorio con reposición es más pequeño que la varianza en el muestreo sistemático.

Este valor de ρ se representará por ρ_2.

d.- Así los valores de ρ que hacen aconsejable utilizar muestreo sistemático se encuentran en el intervalo:

a) $n > 1$

b) $\rho < 0$

c) Si $\rho \geq \frac{-1}{n-1}$ entonces $V\left(\bar{x}_j\right) \geq V(\bar{x})$.

Si, bajo estos límites se calcula la función $\rho = \frac{-1}{n}$, se puede verificar que para valores de n > 10, el coeficiente ρ converge hacia cero y se aproxima a su valor optimo, donde $V\left(\bar{x}_j\right) = V(\bar{x})$.

Todo estos valores del coeficiente de correlación intraclásica se definen sólo en términos de población y tamaño de la muestra. Para una aplicación concreta, una selección sistemática, con una cierta población marco, entonces ρ tomará un valor que será función de el tamaño de muestra, el tamaño de población y de la distribución de la variable. A este valor de ρ nosotros le denominaremos ρ_3.

e.- En el gráfico 4 podemos ver los valores del coeficiente de correlación intraclásico ρ_0, y del coeficiente ρ_1. Así, para valores diferentes de n el valor de ρ se ha calculado con dos hipótesis diferentes y complementarias:

1.- $\rho_1 = \frac{-1}{n}$. Si el valor de ρ_1 obtenido es igual al valor proporcionado por una muestra concreta, ρ_3 entonces $\frac{Vcr}{n} = Vsist$

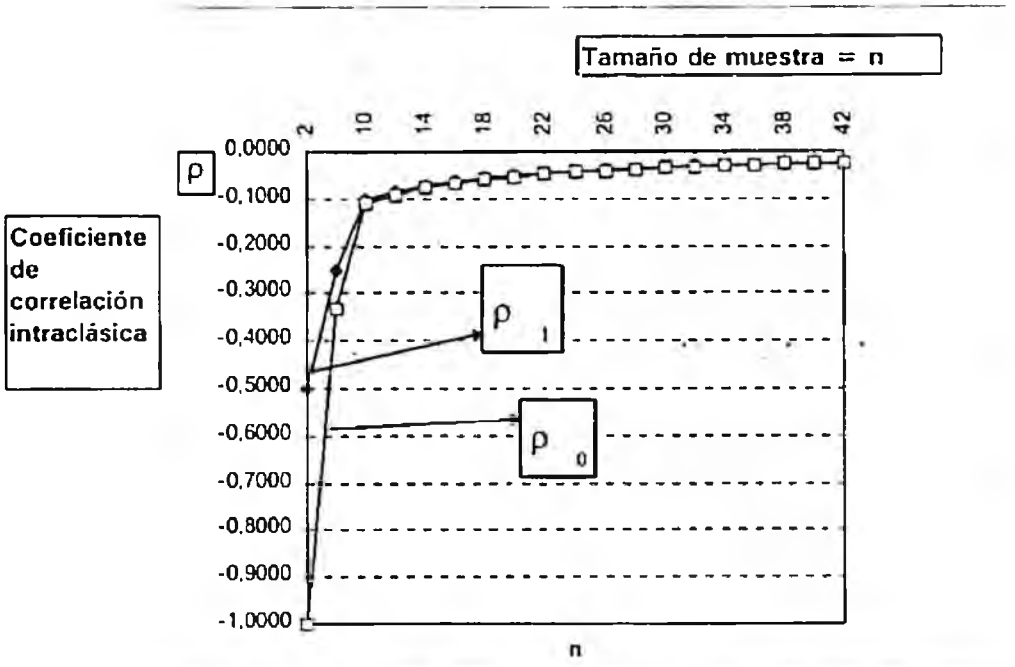
2.- $\rho_0 = \left[\frac{-1}{n-1}\right]$. El coeficiente obtenido en una muestra concreta, ρ_3 no puede ser inferior al valor obtenido con esta

expresión puesto que la varianza no puede ser negativa. En ese caso, $\rho_3 = \left[\frac{-1}{n-1}\right] = \rho_0$, tendríamos para el muestreo sistemático una varianza igual a cero.

Interpretación del gráfico 4:

Si ρ_3 es negativo y se puede despreciar la fracción de muestreo, entonces, si el valor ρ_3 es menor o igual a ρ_1, la varianza del muestreo sistemático se puede aproximar por el cociente de la varianza para el muestreo aleatorio con reposición y el tamaño de muestra.

Gráfico 4.- Límites del coeficiente de correlación intraclásica (ρ_0 , ρ_1) para diferentes tamaños de muestra.



f.- Vamos a suponer que no se puede despreciar la fracción de muestreo y que consideramos un esquema de muestreo sin reposición.

Con un razonamiento similar al utilizado previamente tendremos:

$$\left.\begin{aligned} V(\bar{x}_s) &= \frac{\sigma^2}{n}\left[1+(n-1)\rho\right] \\ V(\bar{x}) &= \frac{N-n}{N-1}\frac{\sigma^2}{n} \end{aligned}\right\}$$

Así,

$$\frac{V(\bar{x})}{\frac{N-n}{N-1}}\left[1+(n-1)\rho\right] = V(\bar{x}_s) = \frac{V(\bar{x})}{n}$$

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

queda

$$\rho = \left[\frac{-N(n-1)}{n(n-1)(N-1)}\right] = \frac{-N}{n(N-1)}\,\frac{-1}{n(1-\frac{1}{N})} = \frac{-1}{n} = \rho_1$$

Si $\frac{1}{N}$ se puede despreciar.

Así, si $\rho_2 = \rho_1$ entonces podríamos usar $\frac{V(\bar{x})}{n}$ para obtener una aproximación de la varianza de muestreo sistemático, independientemente de si se puede o no despreciar el fracción de muestreo.

Conclusión de la sección f:

La aproximación de la varianza en muestreo sistemático con $\frac{Vsr}{n}$ [1] es válido con la única restricción de que $\frac{1}{N}$ se pueda despreciar, independientemente de si se puede o no despreciar la fracción de muestreo.

Con la expresión -1/ n, es posible calcular el valor ρ_1, que establece la condición según la cual se podría aplicar la aproximación propuesta en el caso que se verifique que ρ_3 es menor o igual que ρ_1.

g.- Por último para encontrar un límite, en el caso que no se puede despreciar la fracción de muestreo, más preciso que en el caso donde se despreció esta fracción, donde sólo tenía la condición de no tomar valores negativos, podríamos argumentar que:

$$\left.\begin{aligned} V(\bar{x}_1) &= \frac{\sigma^2}{n}\left[1+(n-1)\rho\right] \\ V(\bar{x}) &= \frac{N-n}{N-1}\frac{\sigma^2}{n} \end{aligned}\right\}$$

[1] Como de costumbre usamos " sr" por las iniciales de " muestreo sin reposición."

$$\left.\begin{aligned}\frac{\sigma^2}{n} &= \frac{V(\bar{x}_s)}{1+(n-1)\rho}\\ \frac{\sigma^2}{n} &= \frac{V(\bar{x})}{\frac{N-n}{N-1}}\end{aligned}\right\} \Rightarrow \frac{V(\bar{x}_s)}{1+(n-1)\rho} = \frac{V(\bar{x})}{\frac{N-n}{N-1}}$$

Y como para que

$$V(\bar{x}) > V(\bar{x}_s) \qquad \frac{V(\bar{x}_s)}{V(\bar{x})} < 1$$

es necesario que $\rho_2 < -\frac{n-1}{(N-1)(n-1)} \leq -\frac{1}{N-1}$

Éste es un límite superior a partir del cual es preferible no aplicar muestreo sistemático frente a muestreo aleatorio simple. Para ilustrar este esquema se puede observar el gráfico 2, deducido de las anteriores expresiones, donde los cálculos se realizan para una fracción de muestreo de 0.1.

7.2. ANÁLISIS GRÁFICO

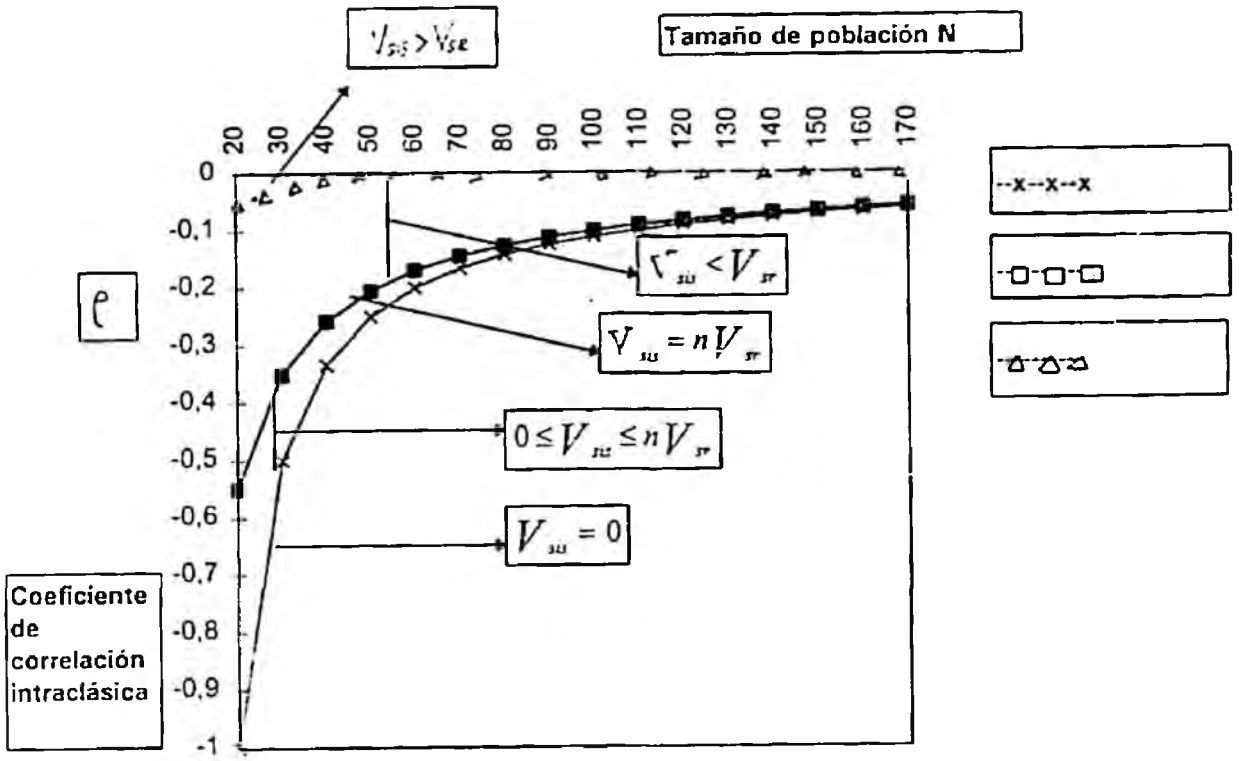
De los límites deducidos en las secciones anteriores es posible llevar a término el siguiente análisis gráfico. Que permite el estudio del coeficiente de correlación intraclásica en función del valor de la varianza del muestreo sistemático y de la varianza del muestreo aleatorio sin reposición, para varios tamaños del población y muestra, fracción igual a 0,1. Así cuando ante un diseño sistemático concretado por un determinado marco poblacional y un determinado tamaño de muestra el coeficiente de correlación intraclásica se podrán observar los siguientes casos: $= \rho_1$

1.- Si el coeficiente ρ_3 calculado es mayor que ρ_2 entonces la varianza del muestreo sistemático será mayor que la varianza del muestreo aleatorio simple sin reposición.

2.- Si ρ_3 esta comprendido entre ρ_2 y ρ_1 entonces la varianza del muestreo sistemático será menor que la varianza del muestreo aleatorio simple sin reposición.

3.- Si ρ_3 esta comprendido entre ρ_1 y ρ_0 entonces la varianza del muestreo sistemático será menor que n veces la varianza del muestreo aleatorio simple sin reposición.

Gráfico 5.-[2] Varianza del estimador para diferentes combinaciones de tamaños de población



[2] V representa la varianza del estimador para cada diferente tipo de muestreo utilizado.

7.3. ILUSTRACIÓN CON DATOS DEL MARCO DE LA ENCUESTA DE POBLACIÓN ACTIVA

Los principales resultados obtenidos[3] para un tamaño de muestra igual a 20 hogares , utilizando la aplicación POSDEM han sido:

$\overline{X} = 3{,}13 \quad \sigma^2 = 2{,}19 \quad Cv_{SR} = 10{,}3\% \quad Cv_{s;r} = 1.7\%$

Para intervalo de muestreo constante

$\rho_{3} = -0.048$

$S_b^2 = 0.21$[4]

$S_w^2 = 2.29$[5]

$\hat{V}_{sis} = 0.01$

$C_v = \frac{\sqrt{V_{sis}}}{\hat{X}} \times 100 = 3.2\ \%$

Para intervalo de muestreo variable

[3] Marco: población referenciada de 400 unidades. Nombre del fichero que contiene los datos Epa400.asc

[4] Cuasivarianza entre muestras.

[5] Cuasivarianza dentro de muestras.

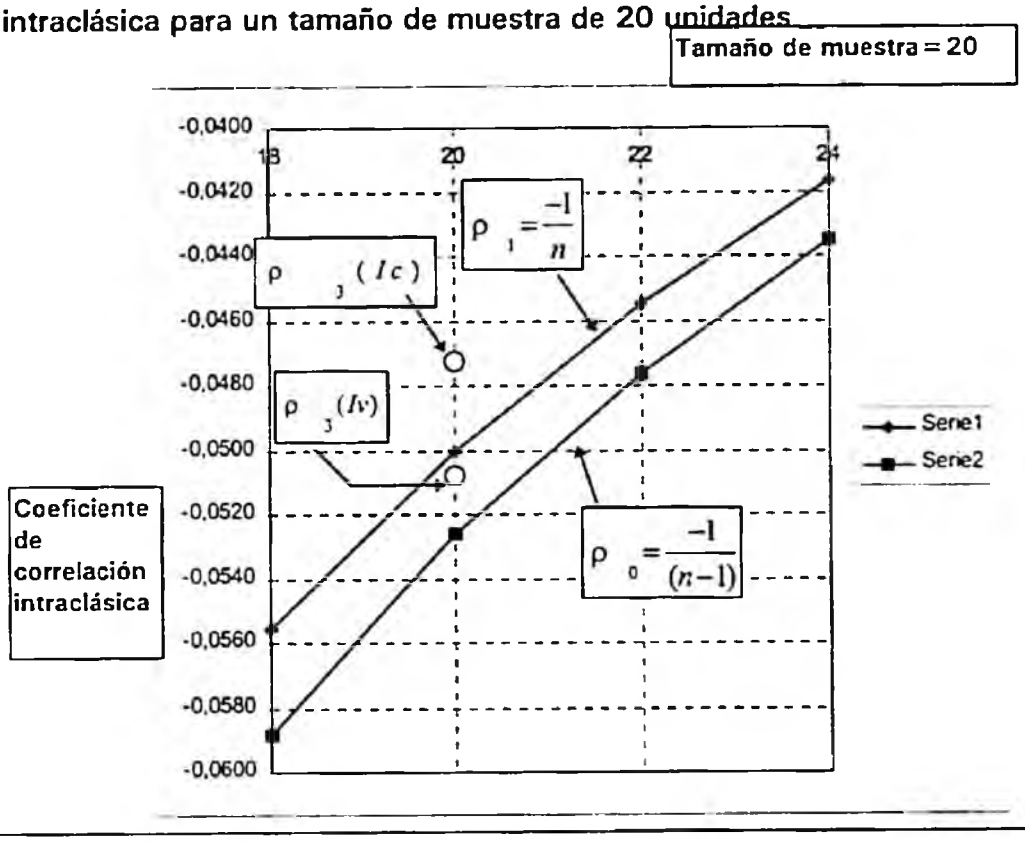
$\rho_{3} = -0.051$

$S_b^2 = 0.054$

$S_w^2 = 2.30$

$V_{sis} = 0.0026$ $\qquad C_v = \frac{\sqrt{V_{su}}}{\bar{X}} \times 100 = 1.6\ \%$

Gráfico 6.- Aplicación al marco de la EPA[6]. Coeficientes de correlación intraclásica para un tamaño de muestra de 20 unidades



[6] EPA: Encuesta de Población Activa.

Aplicación a los datos del marco de la Encuesta de Población Activa.-

Al considerar sobre la misma población un tamaño de muestra de 40 hogares los resultados son:

$\overline{X} = 3{,}13 \quad \sigma^2 = 2{,}19 \quad CV_{mar} = 7{,}09\% \quad CV_{str} = 1{,}02\%$

Para intervalo de muestreo constante

$\rho_3 = -0.023$

$S_b^2 = 0.205$

$S_w^2 = 2.24$

$V_{sis} = 0.0046$

$C_v = \frac{\sqrt{V_{sis}}}{\bar{X}} \times 100 = 2.1\ \%$

Para intervalo de muestreo variable

$\rho_3 = -0.0255519$

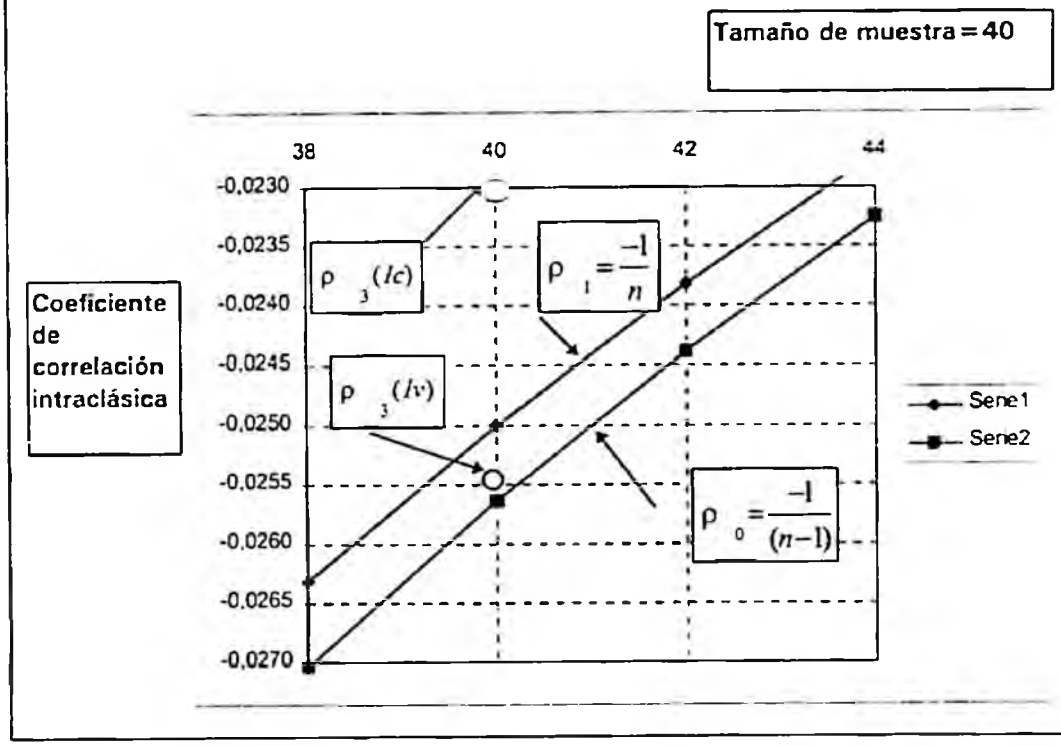
$S_b^2 = 0.011389$

$$S_x^2 = 2.245833$$

$$V_{sis} = 0.0003$$

$$C_x = \frac{\sqrt{V_{sis}}}{\bar{X}} \times 100 = 0.5\ \%$$

Gráfico 7.- Aplicación al marco de la EPA[7]. Coeficientes de correlación intraclásica para un tamaño de muestra de 20 unidades.



[7] EPA: Encuesta de Población Activa.

7.4. ILUSTRACIÓN CON DATOS DEL MARCO DE LA ENCUESTA INDUSTRIAL

Los principales resultados obtenidos[8] para un tamaño de muestra igual a 16 empresas , utilizando la aplicación POSDEM han sido:

$$\overline{X} = 54{,}65 \qquad \sigma^2 = 1585.2 \quad Cv_{sr} = 17.2\% \quad Cv_{str} = 2.3\%$$

Para intervalo de muestreo constante

$$\rho_{3} = -0.061$$

$$S_b^2 = 151.2$$

$$S_w^2 = 1681.8$$

$$V_{sis} = 8.5$$

$$C_v = \frac{\sqrt{V_{sis}}}{\overline{X}} \times 100 = 5.3\ \%$$

[8] Marco: población referenciada de 160 unidades. Nombre del fichero que contiene los datos Ele160.asc

Para intervalo de muestreo variable

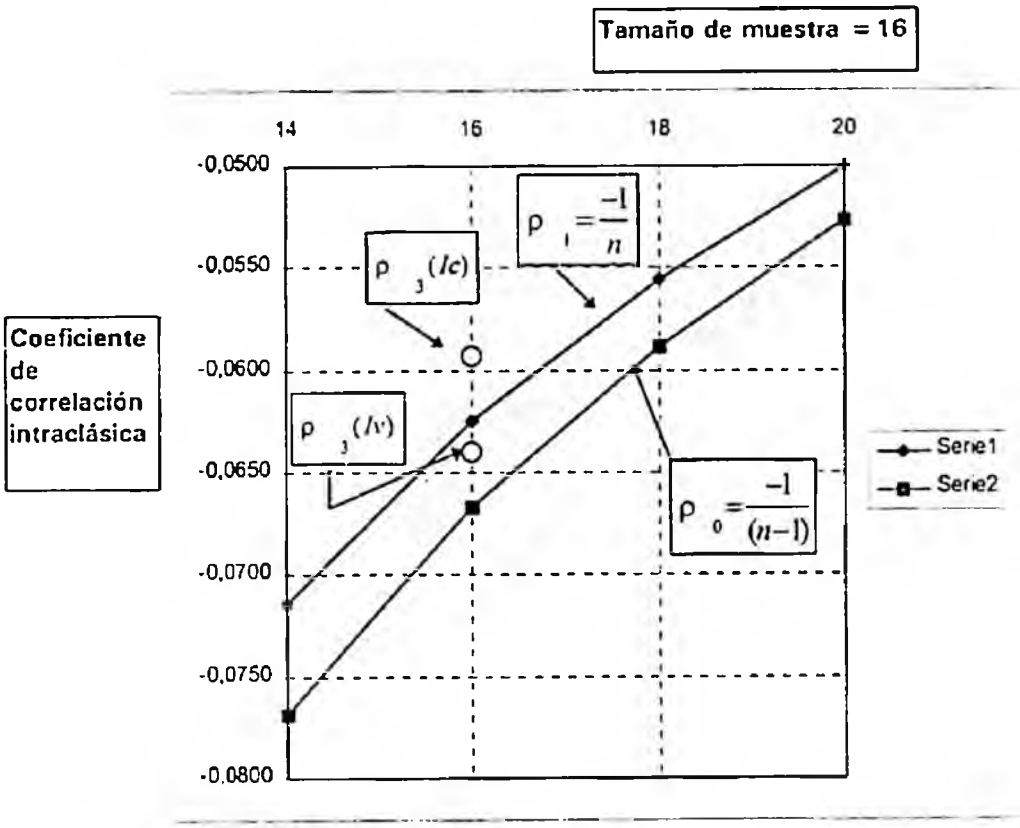
$\rho_{} = -0.065$

$S_b^2 = 31.9$

$S_w^2 = 1689$

$V_{sis} = 1.7$ $\qquad C_v = \frac{\sqrt{V_{sis}}}{\bar{X}} \times 100 = 2.4\ \%$

Gráfico 8.- Aplicación al marco de la EIE[9]. Muestra de 16 empresas.



Aplicación a los datos del marco de la Encuesta Industrial.-

Al aumentar a 32 el tamaño de muestra los resultados son:

$\overline{X} = 54.65 \quad \sigma^2 = 1585.2 \quad Cv_{ir} = 11.51\% \; Cv_{ur} = 0.9\%$

Para intervalo de muestreo constante

[9] EIE: Encuesta Industrial de Empresas.

$\rho_{3} = -0.0308$

$S_b^2 = 86$

$S_w^2 = 1634$

$V_{sis} = 2.15$

$C_v = \frac{\sqrt{V_{sis}}}{\bar{X}} \times 100 = 2.6\ \%$

Para intervalo de muestreo variable

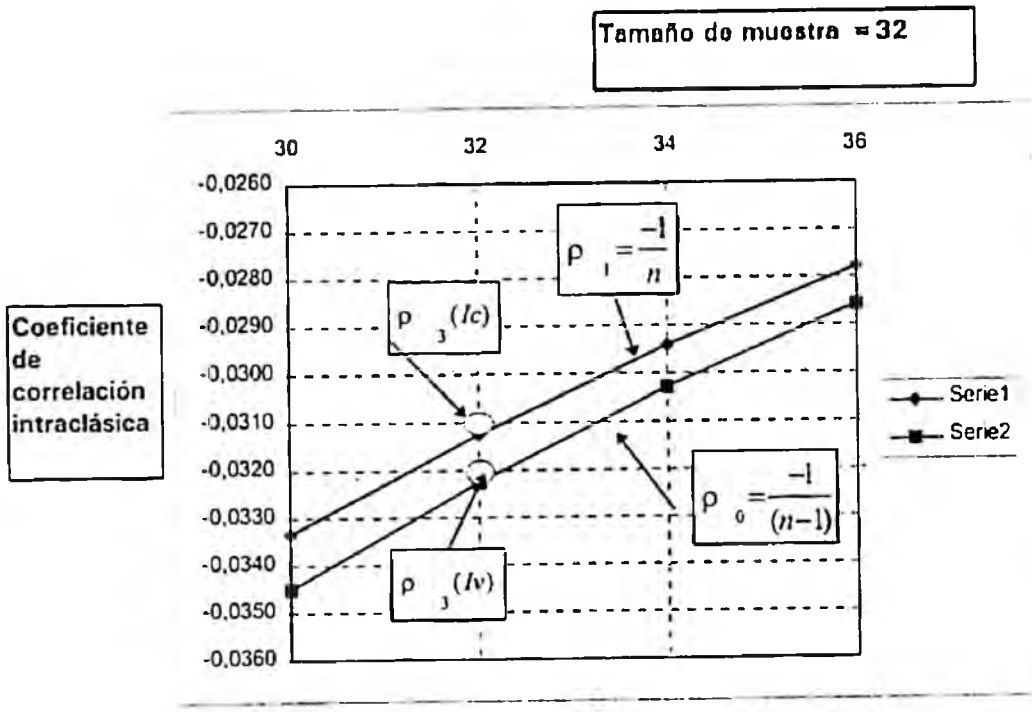
$\rho_{3} = -0.0321$

$S_b^2 = 5.3$

$S_w^2 = 1636$

$V_{sis} = 0.13 \qquad C_v = \frac{\sqrt{V_{sis}}}{\bar{X}} \times 100 = 0.6\ \%$

Gráfico 9.- Aplicación al marco de la EIE[10]. Muestra de 16 empresas.



[10] EIE: Encuesta Industrial de Empresas.

7.5. COMENTARIO DE LOS RESULTADOS

Para observar, en una única tabla, la ganancia o perdida de precisión que se produce al usar un método de muestreo u otro, se puede utilizar el siguiente esquema de relación entre los coeficientes de variación de los estimadores correspondientes a los cuatro métodos analizados.

Aplicación a los datos de la Encuesta de Población Activa. Con las especificaciones, marco: Epa400.asc, N= 400 y n= 20, los resultados obtenidos para el coeficiente de variación expresado en términos relativos han sido los siguientes[11].

Cv (M_sr)= 10% Cv (M_sis_imv)= 1,6%

Cv (M_sis_imc= 3,2% Cv (M_strt_nh=1= 1,7%

[11] No se explican los subíndices por ser evidentes.

De aquí se puede deducir la siguiente tabla:

Tabla 15.- Ganancias y perdidas de precisión para métodos alternativos.

A usar (2) / De usar (1)	M_sr	M_sis_imc	M_sis_imv	M_str_nh = 1
M_sr	0%	-222%	-544%	-506%
M_sis_imc	69%	0%	-100%	-88%
M_sis_imv	84%	50%	0%	6%
M_str_nh = 1	83%	47%	-6%	0%

Se define la ganancia (valores positivos) o perdida (valores negativos) de precisión de usar el método (1) a usar el método (2), por la expresión:

$$G_{1/2} = (1 - \frac{C_{v1}}{C_{v2}})100$$

Podemos interpretar estos valores de la siguiente forma, si utilizamos un método u otro, se produce una ganancia de precisión (+) o perdida (-), que puede expresarse en porcentaje del valor de la precisión del método que se pretende sustituir. Así, por ejemplo, de usar el método de muestreo sistemático con intervalo variable, que tiene una

precisión del 1,6% a usar muestreo estratificado con una unidad por estrato, que tiene una precisión del 1,7%, se produce una ganancia en precisión del 6% con respecto al mismo 1,7%.

El caso más interesante será aquel en el que todo los valores de la fila, correspondiente a un cierto método, sean todos positivos. Significará que se produce una ganancia en precisión en relación con cualquier otro método considerado. En este ejemplo se verifica que el método que mejora en precisión, a los restantes, es el método de muestreo sistemático con intervalo variable.

Para la población considerada se puede observar como la ganancia en precisión al usar muestreo sistemático con intervalo variable frente al intervalo constante supone una reducción del error relativo de 3,2% a 1,6%, lo que parece un resultado prometedor. También es otro resultado interesante que el método con intervalo variable es más preciso que el método estratificado con una unidad por estrato, que supone una ganancia de precisión del 6%.

Se puede observar también como al aumentar el tamaño de muestra 40 hogares la ganancia en precisión del método sistemático con intervalo variable es más importante aun que para $n=20$. Sólo en un caso de los estudiados para la Encuesta Industrial y tamaño de muestra igual a 16 empresas el muestreo estratificado con una unidad por estrato supera en precisión al sistemático con intervalo variable.

CAPÍTULO 8

MODELOS DE SUPERPOBLACIÓN

8.1. RESULTADOS A TRAVÉS DE MODELOS DE SUPERPOBLACIÓN

Utilizando la aplicación POSDEM vamos a modelar el comportamiento de una población para estudiar la varianza del estimador para diferentes métodos respecto del modelo y su propia variabilidad. En aquellos métodos de muestreo que son insesgados, centrados y de Yates, se utilizará el error cuadrático medio en lugar de la varianza del estimador.

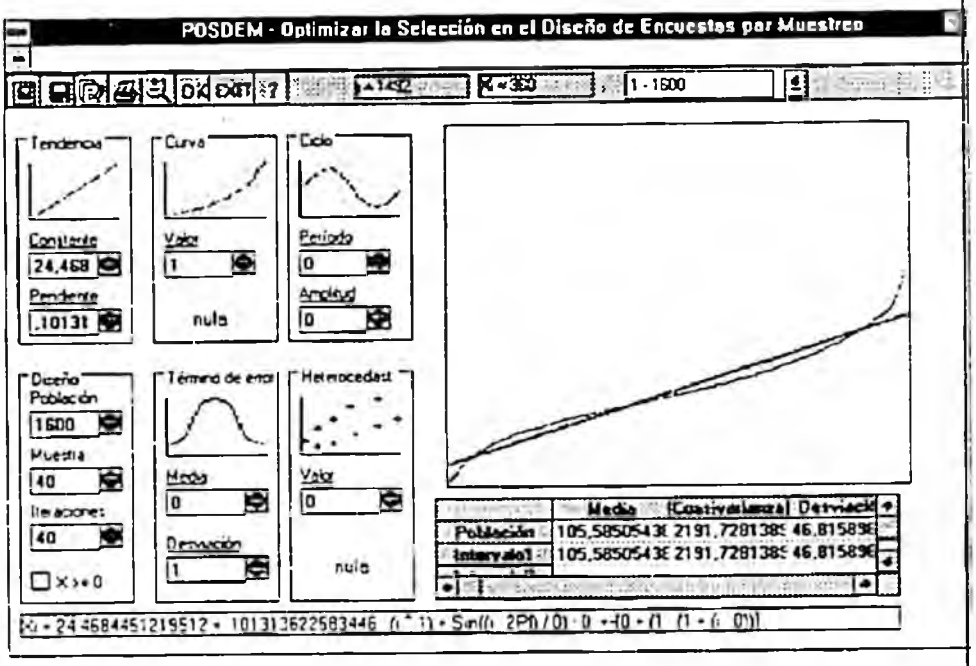
Este enfoque surge de la necesidad de inferir resultados más allá de lo que representa el análisis de una única población natural, consecuencia de una determinada realización.

Así consideramos la "población marco" compuesta por 1600 secciones censales correspondientes al País Vasco[1]. Para esta población disponemos de abundante información censal, casi 3000 datos por sección. Para el estudio que nos ocupa podemos ordenar las secciones, que serán nuestras unidades de muestreo, por la variable "número de parados en la sección". La representación gráfica de los datos poblacionales y su correspondiente ajuste se puede observar en el gráfico 1. En él se ha incluido una pantalla que

[1] Datos procedentes del CD ROM "Los municipios CERCA. España. Censos 1989-1991" Publicado por el INE

muestra el "simulador de estructuras" utilizado por la aplicación POSDEM.

Gráfico 10.- Simulador de estructuras de población en POSDEM. Determinación de modelos de superpoblación.



Los conceptos que aparecen en esta pantalla hacen referencia a las siguientes cuestiones.

- Tendencia: pone de manifiesto los términos obtenidos por el programa en una regresión lineal. Admite otros procedimientos para el término constante y la pendiente.
- Curva: se refiere al grado del polinomio que queremos ajustar. En caso de que sea mayor que uno para ver los valores obtenidos es necesario observar la ecuación que aparece en la última línea.
- Ciclo: permite definir los parámetros que simulan el periodo y la amplitud en el caso de existencia de un movimiento cíclico.
- Diseño de población: se refiere a las características tamaño de población, de muestra y al cociente entre ambas, que define el número de grupos que se van a utilizar en la formación del espacio muestral en los diseños sistemáticos.
- Término de error: se distribuye según una normal de parámetros media y desviación.
- Heterocedásticidad: permite introducir un comportamiento de heterocedásticidad en la distribución del término de error.

En la tabla tenemos los cálculos obtenidos al aplicar a la población natural dei País Vasco los diferentes métodos de selección de muestras.

Tabla 16.- Población marco de 1600 secciones censales del País Vasco.[2] Variable de ordenación número de parados.

Datos Población N = 1600 n = 40 $\bar{x} = 10557$ $s^2_{n-1} = = 2344.32$	Varianza del Estimador media[3]
Método	$V(\bar{\bar{x}})$
Sin reposición	57.14
Intervalo constante	4.29
Intervalo variable	.9263
Modificado	.5300
Estratificado con una unidad	.4823
Intervalo variable II	.4494
Centrado con intervalo variable II	.3255
Equilibrado	.2251
Corregido de Yates	.1318
Centrado de Madow	.0740[4]
Centrado con intervalo variable	.0677

Si únicamente tuviésemos en cuenta esta información el método de selección elegido, de forma que se minimice la varianza

[2] Tabla de elaboración propia, utilizando el programa POSDEM y con datos procedentes del Cd.ROM "Municipios.CERCA. Censos 1989-1991" Fuente INE.

[3] En los métodos sesgados, centrados y Yates, se utiliza el error cuadrático medio en lugar de la varianza.

[4] Este resultado y el siguiente están correctamente calculados. A pesar de ello se observo que toman un valor extremadamente pequeño. Esto da lugar a considerar el método centrado como errático en ciertas ocasiones. Desde la óptica de los modelos de superpoblación, esto tiene su justificación por la inestabilidad puesta de manifiesto por los elevados valores que suele tomar, respecto de otros métodos, la varianza respecto del modelo del error cuadrático medio del estimador. Que hemos observado también que presenta una relación positiva con el aumento de la variabilidad aleatoria del modelo.

para ese tamaño de muestras, sería el método centrado con intervalo variable, que presenta la menor varianza, un valor de 0.0677.

Ahora podemos, mediante el módulo de "simulación de estructuras" de la aplicación POSDEM[5], definir un modelo de superpoblación que explique esta población marco considerada mediante la expresión:

$$x_u = \alpha + \beta_u + e_u$$

donde *u* representa las unidades de la población y en este caso toma los valores de 1 a 1600.

α y β son los parámetros de la recta calculados inicialmente por mínimos cuadrados, pero que pueden ser definidos con otros procedimientos.

e_u es el término de error aleatorio que en este modelo se ha definido con los valores

$$E_m(e_u) = 0 \; ; \; E_m(e_u^2) = \sigma^2 \; ; \; E_m(e_u \; e_v) = 0$$

el operador E_m denota la esperanza respecto del modelo.

Así:

$$x_u = 24.4 + 0.10u + e_u \text{ donde } \sigma^2 = 1$$

[5] Con la utilidad de POSDEM es posible definir modelos no lineales más complejos. Permite utilizar modelos diferentes por tramos de población. La complejidad de los modelos que es posible definir se puede deducir de la fórmula que aparece en la ultima linea del gráfico 1.

Con carácter general, para este tipo de modelo tenemos para:

1) Muestreo sistemático:

$$E_m V_p (\bar{x}_{sis}) = \beta^2 (k^2 - 1)/12 + \bar{\sigma}^2$$

con $\bar{\sigma}^2 = \sigma^2 (k - 1)/nk$

El primer componente es la varianza debida a la tendencia lineal; el segundo término es la debida al error aleatorio.

2) Muestreo aleatorio:

$$E_m V_p (\bar{x}_{SR}) = \beta^2 (k - 1) (nk + 1)/12 + \bar{\sigma}^2$$

3) Muestreo estratificado con una unidad por estrato: Se asume que la población consiste en *n* estratos formados por los conjuntos de unidades $\{1...k\}$ $\{k+1,...2k\}$,...$\{(n-1)k+1,...,nk\}$. Una muestra aleatoria se toma de cada estrato.

$$E_m V_p (\bar{x}_{strat}) = \beta^2 (k^2 - 1)/12n + \bar{\sigma}^2$$

Se puede comprobar que:

$$E_m V_p (\bar{x}_{strat}) \leq E_m V_p (\bar{x}_{sis}) \leq E_m V_p (\bar{x}_{SR})$$

En resumen para el modelo propuesto:

$x_u = 24.4 + 0.10u + e_u$ donde $\sigma^2 = 1$

Tendríamos el siguiente resultado teórico:

$E_m V_p (\bar{x}_{sis}) = 1.3569 \quad E_m V_p (\bar{x}_{strat}) = 0.0577\ E_m V_p (\bar{x}_{SR}) = 53.38$

En nuestra aplicación *con los datos del país vasco* encontramos, que para este modelo, los valores que se obtienen con la aplicación POSDEM coinciden con los que se obtendrían de aplicar los desarrollos teóricos anteriores. Estos valores son:

Tabla 17.- Modelo con tendencia (I). Variable de ordenación número parados.

Método	EmVp( )[6]	VmVp( )	CvmVp%	EmVp+2*raiz(Vp)
Muestreo sin reposición (S.R.)	53.44073	0.00299	0.10%	53.55018
Sistemático intervalo cte. (SIC)	1.40023	0.00278	3.77%	1.50575
Sistemático centrado int. var. 2 (CLSIV2)	0.02801	0.00089	106.42%[7]	0.08761
Sistemático centrado de Madow CLS)	0.02611	0.00060	93.52%	0.07496
Muestreo estratificado con una unidad	0.05886	0.00000	2.43%	0.06173
Sistemático centrado int. var. 1 (CLSIV1)	0.02033	0.00034	91.22%	0.05741
Sistemático intervalo var. 2 (SIV2)	0.02484	0.00004	24.36%	0.03695
Sistemático corrección de Yates (CY)	0.02509	0.00003	23.57%	0.03691
Sistemático intervalo var. 1 (SIV1)	0.02427	0.00003	23.96%	0.03589
Sistemático modificado (MSS)	0.02452	0.00003	22.16%	0.03539
Sistemático equilibrado (BSS)	0.02505	0.00003	20.22%	0.03519

Se ha repetido el proceso de simulación de la población natural un número elevado de veces, que se ha fijado en cincuenta. Si bien no existe inconveniente en aumentar este número, no obstante

[6] EmVp() representa la esperanza respecto del modelo de la varianza del estimador
[7] Ver nota a pie página de la tabla 1.

se puede observar que los valores se estabilizan suficientemente al pasar de 30 a 40 simulaciones. Los datos obtenidos con el procedimiento descrito se han puesto, para los diez métodos considerados, en la tabla 2. Sin tener en cuenta más consideraciones tendríamos que elegir, en base a un modelo de superpoblación, el método de menor varianza esperada. Por tanto elegiríamos el método centrado con intervalo variable que presenta un valor esperado del error de 0.02033. No obstante en este punto disponemos de información relativa a cómo oscila este valor esperado respecto del modelo. Esto es, disponemos de la información contenida en la segunda columna de la misma tabla, concerniente a la varianza sobre el modelo. Así, podemos establecer intervalos con una determinada confianza de que entre sus límites se encuentre el error de muestreo. Con esta nueva información, podemos elegir como mejor aquel método que presente un menor valor del indicador: "Em Vp+ 2*raíz(Vp). Este indicador de la cota superior del error proporciona información no sólo de cuál va a ser el error esperado sino también de su estabilidad. Destacar aquí el que los métodos centrados presentan un error esperado con una gran variabilidad del mismo. En otro extremo, el muestreo estratificado con una unidad por estrato

destaca a su vez por su elevada estabilidad en el diseño. Siguiendo con el mismo razonamiento elegiríamos el método sistemático equilibrado que presenta un valor de 0.03519. Se puede observar que todos los métodos considerados, a excepción de los dos primeros, presentan un error muy similar que impide discriminar entre ellos con garantías de estar eligiendo el mejor. Destaca esta situación con la situación original, referida a la población natural, donde se observaba un abanico de diferencias entre los métodos que permitía declarar unos como preferibles a otros. Esto puede ser debido a la simplicidad del modelo que se ha considerado que no aprovecha o no explica, de forma conveniente, la información del marco.

Así, podemos comprobar gráficamente que al sustituir la población por una tendencia lineal se pierde la información contenida en los extremos. Esto nos lleva a plantear un segundo modelo similar al anterior, pero algo más complejo, en el sentido de que en lugar de utilizar una única ecuación, utiliza tres. Una primera para describir el comportamiento de las 200 primeras secciones. Una segunda ecuación que describe el comportamiento de las secciones 201 a 1530 y una última ecuación que permite modelar el último tramo de población.

Los resultados del segundo modelo se encuentran en la tabla de la tabla y permiten elegir, como mejor método, el método equilibrado que es el que presenta un menor valor en la columna

correspondiente al indicador: EmVp + 2*raíz(Vp). Aunque tenemos una situación similar a la anterior, ahora se puede discriminar más los métodos que proporcionan un resultado diferente.

Tabla 18.- Modelo con tendencia (II). Variable de ordenación número de parados.

Método	EmVp( )	VmVp( )	CvmVp	EmVp + 2*raíz(Vp)
uestreo sin reposición (S.R.)	56.29427	0.00329	0.00102	56.40906
Sistemático intervalo cte. (SIC)	3.77968	0.00703	0.02219	3.94741
Sistemático centrado int. var. 2 (CLSIV2)	0.49612	0.02561	0.32259	0.81621
Sistemático intervalo var. 1 (SIV1)	0.64984	0.00221	0.07227	0.74376
Sistemático intervalo var. 2 (SIV2)	0.42913	0.00169	0.09588	0.51142
Muestreo estratificado con una unidad	0.34307	0.00002	0.01161	0.35104
Sistemático modificado (MSS)	0.24170	0.00052	0.09446	0.28736
Sistemático corrección de Yates (CY)	0.10897	0.00017	0.11997	0.13512
Sistemático centrado de Madow (CLS)	0.03140	0.00109	1.05042	0.09738
Sistemático centrado int. var. 1 (CLSIV1)	0.02832	0.00079	0.98994	0.08439
Sistemático equilibrado (BSS)	0.02977	0.00004	0.20596	0.04204

Podemos aumentar el número de tramos para modelar esta población o bien utilizar un modelo con ajuste por polinomios ortogonales de grado superior a dos. Incluso se podría, si fuese necesario para modelar fielmente la población, el combinar el establecimiento de tramos con el ajuste por polinomios de grado entre uno y cinco. Se ha optado por realizar un ajuste de un polinomio de grado cinco, sin definir tramos, y los resultados, bajo el título de modelo III, permiten limitar a cinco el número de métodos aconsejables para esta población marco. Estos resultados se muestran en la tabla 4. Al tomar una decisión de cual es el método preferible, sin tener en cuenta más aspectos que los considerados,

probablemente seleccionaríamos el método de muestreo equilibrado por los resultados del modelo, en especial por su estabilidad y porqué en la población marco original presentaba un resultado intermedio pero situado entre los cinco mejores. Se excluyen en este análisis los métodos centrados por presentar un elevado grado de inestabilidad, reflejado en coeficientes de variación muy altos, no obstante se resalta que presentan buenos resultados en cuanto al valor esperado del error cuadrático medio. En el caso de que el método equilibrado o el de correcciones de Yates no hubiesen proporcionado buenos resultados en esta población hubiese sido aconsejable optar por el método centrado. En caso de duda por similitud de resultados sería necesario atender al comportamiento del método para estimar el error en base a los datos de una sola muestra.

Tabla 19.- Modelo con tendencia (III). Variable de ordenación número de parados.

Método	EmVp( )	VmVp( )	CvmVp	EmVp + 2*raíz(Vp)
Muestreo sin reposición (S.R.)	56.94545	0.00349	0.00104	57.06363
Sistemático intervalo cte. (SIC)	3.55784	0.00747	0.02430	3.73072
Sistemático intervalo var. 1 (SIV1)	0.51325	0.00087	0.05742	0.57220
Sistemático modificado (MSS)	0.24891	0.00052	0.09153	0.29448
Muestreo estratificado con una unidad	0.21484	0.00001	0.01449	0.22106
Sistemático centrado int. var. 2 (CLSIV2)	0.09192	0.00328	0.62348	0.20654
Sistemático intervalo var. 2 (SIV2)	0.13525	0.00039	0.14599	0.17474
Sistemático corrección de Yates (CY)	0.06979	0.00013	0.16071	0.09222
Sistemático centrado de Madow (CLS)	0.02755	0.00085	1.05562	0.08570
Sistemático centrado int. var. 1 (CLSIV1)	0.02442	0.00040	0.81510	0.06423
Sistemático equilibrado (BSS)	0.03066	0.00006	0.24710	0.04582

Otro análisis similar puede realizarse cuando la variable de ordenación es el "número de habitantes de la sección". En este caso, dada la estructura de la población natural, se ha utilizado un ajuste polinómico de grado cinco con desviación típica del término aleatorio igual a 10 y coeficiente de heterocedastícidad igual a 0.001. Así el término aleatorio del modelo queda: +/- (0+(10*(1+(i*0.001)))). Los resultados se muestran en la tabla 5 y 6. Estos cálculos presentan un resultado más fácilmente interpretable como óptimo en el método de muestreo estratificado con una unidad por estrato. En este caso contrasta esta información proporcionada por los modelos frente a la proporcionada por la población marco original. El decidir optar por un método u otro dependerá del grado de fiabilidad que nos proporcione el marco o, en la otra cara de la moneda, del grado de seguridad que asignemos al modelo. Por ejemplo, con una población de estas características, escaso grado de correlación entre las variables población y paro, en torno al 0.6, si nos encontramos en un momento próximo al de la realización del censo y tenemos confianza en que su acuracidad es elevada y pensamos que son variables con un componente estructural, a pesar del resultado del modelo, podría ser conveniente seleccionar el método equilibrado, siguiendo un razonamiento similar al del apartado anterior. Si, por el contrario, pensamos que la variable de estudio es coyuntural y que el dato de población se puede actualizar con una elevada exactitud y que se trata de una variable relacionada con el conjunto de la encuesta entonces parece más aconsejable seguir el consejo de los modelos y optar por un método estratificado con una unidad.

Tabla 20.- Población marco de 1600 secciones censales del País Vasco.[8] Variable de ordenación número de habitantes de la sección.

Datos Población N = 1600 n = 40 $\bar{x} = 10557$ $S^2_{n-1} = = 2344.32$	Varianza del Estimador media
Método	$V(\bar{\bar{x}})$
Muestreo sin reposición (S.R.)	57.14299
Sistemático centrado int. var. 2 (CLSIV2)	28.25571
Muestreo estratificado con una unidad	22.70511
Sistemático intervalo var. 2 (SIV2)	17.74366
Sistemático intervalo var. 1 (SIV1)	16.87094
Sistemático modificado (MSS)	16.12801
Sistemático corrección de Yates (CY)	13.94691
Sistemático intervalo cte. (SIC)	13.91532
Sistemático equilibrado (BSS)	12.68813
Sistemático centrado de Madow CLS)	1.20296[9]
Sistemático centrado int. var. 1 (CLSIV1)	1.11296

[8] Tabla de elaboración propia, utilizando el programa POSDEM y con datos procedentes del Cd.ROM "Municipios.CERCA. Censos 1989-1991" Fuente INE.

[9] Ver nota al pie de página de la tabla 1.

Tabla 21.- Modelo con tendencia (IV). Variable de ordenación número de habitantes de la sección.

Método	EmVp( )	VmVp( )	CvmVp	EmVp+2*raíz(Vp)
uestreo sin reposición (S.R.)	43.09524	0.68545	0.01921	44.7510
Sistemático centrado int. var. 2 (CLSIV2)	7.21433	60.69672	1.07991	22.79595
Sistemático centrado de Madow (CLS)	7.66125	47.91670	0.90353	21.50563
Sistemático centrado int. var. 1 (CLSIV1)	5.49311	43.04323	1.19436	18.61458
Sistemático intervalo cte. (SIC)	9.53738	3.90121	0.20710	13.48767
Sistemático intervalo var. 1 (SIV1)	8.95667	3.53659	0.20996	12.71784
Sistemático intervalo var. 2 (SIV2)	8.33457	3.71568	0.23128	12.18979
Sistemático equilibrado (BSS)	8.00487	2.40946	0.19391	11.10936
Sistemático corrección de Yates (CY)	7.85744	2.38312	0.19647	10.94491
Sistemático modificado (MSS)	7.75140	2.41160	0.20034	10.85726
Muestreo estratificado con una unidad	8.44365	0.11238	0.03970	9.11412

CAPÍTULO 9

MUESTREO CON PROBABILIDADES DESIGUALES (I)

9.1. MUESTREO CON PROBABILIDADES PROPORCIONALES AL TAMAÑO.

El método de selección condiciona el método de estimación. En las secciones anteriores todas las estimaciones se basaban en que todas las unidades de la población tenían la misma probabilidad de ser seleccionada. Este esquema general se denomina con probabilidades iguales y tiene su representación en que el estimador no precisa hacer un tratamiento diferenciado para cada unidad que pertenece a la muestra. Así por ejemplo si de una población (1,2,3) se obtiene con reposición y probabilidades iguales una muestra de tamaño 2 igual a (2,3) el estimador para el total es:

$$\hat{x} = N\bar{x} = N\sum_{i=1}^{n}\frac{x_i}{n} = \sum_{i}^{n} x_i \frac{1}{(\frac{n}{N})} = \sum_{i=1}^{n} x_i F_e$$

$$con \quad F_e = \frac{1}{\frac{n}{N}}$$

Donde Fe = factor de elevación igual al inverso de la probabilidad que tiene la unidad i de pertenecer a la muestra, que permanece constante para todas las unidades de la misma.

Así, $\bar{x} = 2*(3/2)+3*(3/2) = 7.5$

Es fácil comprobar que el valor poblacional del total es 6.

Ahora bien es conveniente, puesto que reduce el error de muestreo utilizar información auxiliar disponible para definir probabilidades desiguales de selección para cada unidad. En este caso el factor de elevación, el inverso de la probabilidad de que una unidad pertenezca a la muestra, será diferente para cada unidad. En el ejemplo que estamos considerando se podría introducir una variable Mi de forma que el esquema de población considerada pasaría a ser.

$U_i = 1, 2, 3$

$X_i = 1,2,3$

$M_i = 4,6,10$

donde la suma de $M_i = 20$

Si la muestra se selecciona ahora con probabilidades proporcionales a M_i y si se hubiera seleccionado la misma muestra, utilizando un procedimiento con reposición, el cálculo del estimador para el total es:

$$\hat{x} = \sum_{i=1}^{n} x_i \; F_e = \sum_{i=1}^{n} x_i \; \frac{i}{\Pi_i}$$

$$donde \; \Pi_i = n p_i = n\frac{M_i}{M}$$

Así, $\hat{x} = 2/(2*6/20)+3/(2*10/20)=2/.6+3/.1=6.33$

Que se aproxima, bastante más que en el caso anterior, al valor poblacional del parámetro que se desea estimar.

Es fácil comprobar que si hacemos las probabilidades de selección proporcionales a los valores de la variable todas las muestras posibles proporcionan un estimador igual al valor poblacional.

9.2. MUESTREO CON REPOSICIÓN. ESQUEMA DE HANSEN- HURWITZ

Esta es la primera aproximación al problema de la selección y estimación con probabilidades desiguales en 1943. Estos autores definieron un esquema de muestreo que incorporaba probabilidades proporcionales al tamaño, un método de estimación y de evaluación del error debido al muestreo. Este enfoque equivale en un esquema de urna a introducir tantas bolas para cada unidad de la población como unidades tenga el valor del tamaño de esa unidad. Una vez seleccionada una bola de la urna se obtendrá la unidad asociada con dicha unidad y se devolverá a la urna la bola. De esta forma el esquema de la urna no sufre variación al seleccionar las unidades que formarán la muestra. Con este método la probabilidad de selección de una unidad permanece invariable independientemente del número de bolas que se hayan seleccionado. Y, por tanto, la probabilidad de que una unidad pertenezca a una muestra de tamaño *n* es:

$$P[u_i \in \underline{x}] = n\ p_i$$

El estimador para el total es:

$$\hat{X}_{HH} = \frac{1}{n} \sum_{i}^{n} \frac{X_i}{P_i}$$

y la varianza

$$V(\hat{X}_{HH}) = \frac{1}{n} \sum_{i}^{N} P_i \left(\frac{X_i}{P_i} - X\right)^2$$

Para estimar la varianza se utiliza la fórmula

$$\hat{V}(\hat{X}_{HH}) = \frac{1}{n(n-1)} \sum_{i}^{n} \left(\frac{X_i}{P_i} - \hat{X}_{HH}\right)^2$$

Se trata de estimadores insesgados por lo que

$$E(\hat{X}_{HH}) = X \qquad E(\hat{V}(\hat{X}_{HH})) = V(\hat{X}_{HH})$$

Vamos a desarrollar un ejemplo con la ayuda de una hoja de cálculo. Consideramos una población con valores Xi= 1, 3, 4 y tamaños Mi= 3, 5, 7 respectivamente. Con el método de selección descrito para el esquema con reposición y probabilidades proporcionales a los tamaños podemos formar el espacio muestral y calcular las estimaciones para el total, su varianza y la varianza de la varianza:

Tabla 22.- Hoja de cálculo para un esquema de Hansen-Hurwitz.

HH*	1	2	3	4	5*	6	7	8	9*	10	11
	X1	X2	M1	M2	Xhh	P(x)	E(xhh)	Var(hh)	V(xhhm)	E(v)	V(V)
1.-	1	1	3	3	5,00	0,0400	0,2000	0,3600	0,0000	0,0000	0,0522
2.-	1	3	3	5	7,00	0,0667	0,4667	0,0667	4,0000	0,2667	0,5442
3.-	1	4	3	7	6,79	0,0933	0,6333	0,1376	3,1888	0,2976	0,3907
4.-	3	1	5	3	7,00	0,0667	0,4667	0,0667	4,0000	0,2667	0,5442
5.-	3	3	5	5	9,00	0,1111	1,0000	0,1111	0,0000	0,0000	0,1451
6.-	3	4	5	7	8,79	0,1556	1,3667	0,0960	0,0459	0,0071	0,1872
7.-	4	1	7	3	6,79	0,0933	0,6333	0,1376	3,1888	0,2976	0,3907
8.-	4	3	7	5	8,79	0,1556	1,3667	0,0960	0,0459	0,0071	0,1872
9.-	4	4	7	7	8,57	0,2178	1,8667	0,0711	0,0000	0,0000	0,2844
Totales =					67,71	1,0000	8,0000	1,1429		1,1429	2,7259

La interpretación de las columnas es la siguiente:

En las dos primeras aparecen las unidades seleccionadas para muestras de tamaño n=2. En la tercera y cuarta tenemos el valor de Mi asociado con cada unidad muestral. En la quinta columna aparece el estimador para el total de Hansen-Hurwitz, que para la primera muestra seleccionada (1,1) proporciona el valor 5.

$$\hat{X}_{HH} = \sum \frac{x_i}{n P_i} = \frac{1}{2 * \frac{3}{15}} + \frac{1}{2 * \frac{3}{15}} = 5$$

[1] Muestras que es posible obtener con el procedimiento de HH

[2] Estimador del total para cada muestra.

[3] Estimador de la varianza para cada muestra.

En la sexta columna tenemos la probabilidad de seleccionar esa muestra

$$P(\underline{x}) = \frac{M_1}{M}\frac{M_2}{M} = \frac{3}{15}\frac{3}{15} = 0.04$$

En la séptima columna, y en base a las dos anteriores podemos calcular la esperanza del estimador para cada muestra

$$E(\hat{x}_{HH}) = \hat{X}_{HH}\, P(\underline{x}) = 0.2$$

con suma igual a seis, el valor poblacional

En la octava columna podemos calcular con un procedimiento similar al anterior la varianza del estimador

$V(\hat{X}_{HH}) = (\hat{X}_{HH} - E\hat{X}_{HH})^2\, P(\underline{x}) = 0.36$ que al sumar toda la columna proporciona el valor de 1.142

En la novena columna se obtiene la varianza del estimador para cada muestra, en base a la formula

$$\hat{V}(X_{Hhm}) = \frac{1}{2}\sum\left(\frac{X_i}{P_i} - \hat{x}_{HH}\right)^2 =$$

$$= \frac{1}{2}\left[\left(\frac{1}{\frac{3}{15}} - 5\right)^2 + \left(\frac{1}{\frac{3}{15}} - 5\right)^2\right] = 0$$

para las demás muestras sucesivamente

En las columnas diez y once podemos obtener, teniendo en cuenta la función de probabilidad y las estimaciones, la esperanza y la varianza de la varianza

$$E(\hat{V}) = V(\bar{X}_{Hhm})\,P(\underline{x}) = 0 \_ \Sigma = 1.14^{4}$$

$$V(v) = (\hat{V}(X_{Hhm}) - E(v))^{2}\,P(\underline{x}) = 0.05 \_ \Sigma = 2.72$$

En resumen, en este ejemplo tendremos que la varianza del estimador del total es 1,1429 y la varianza del estimador de la varianza es 2,7259. Estos datos pueden obtenerse, con el mismo procedimiento de cálculo mediante la aplicación POSDEM.

[4] Por la expresión _Sumatorio entendemos la suma de la columna.

9.3. MUESTREO SIN REPOSICIÓN. ESQUEMA DE HORVITZ-THOMPSON

Con este esquema se evita tener la misma unidad varias veces en la muestra. Para ello, en un esquema de urna, una vez que se selecciona una bola se extraen de la urna todas las correspondientes a esa determinada unidad. Tiene la ventaja de presentar mínima varianza pero con el inconveniente de utilizar las probabilidades Π_{ij} de difícil manejo. Brewer ha presentado un método válido solamente para tamaños de muestra igual a dos y valores de P_i distintos a un medio, no obstante el cálculo sigue siendo extremadamente complicado. Así el principal inconveniente de este método está en la distorsión producida en las probabilidades de selección que tiene varias consecuencias : a) dificulta los cálculos siendo conveniente fijar tamaños de muestra iguales a dos, a pesar de lo cual los cálculos siguen siendo muy complejos. b) En los estimadores propuestos para el error de muestreo puede darse el caso de varianzas estimadas negativas.

En esta aplicación vamos a utilizar la selección propuesta por Brewer (1983), de forma que el estimador del total será:

$$\hat{X}_{HT} = \sum_{i}^{\dot{s}} \frac{X_i}{\Pi_i} \qquad con \quad E[\hat{X}_{HT}] = X$$

La varianza del estimador es

$$V(\hat{X}_{HT}) = \sum_{i}^{N}\sum_{j>i}(\pi_i\pi_j - \pi_{ij})\left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_j}\right)^2$$

Y un estimador insesgado de la varianza viene dado por

$$\hat{V}(\hat{X}) = \sum_{i}^{n}\sum_{j>i}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_j}\right)^2$$

Donde

$\pi_i = n\,P_i$

$$\pi_{ij} = \frac{2\,P_i P_j}{D}\,\frac{[1-(P_i+P_j)]}{(1-2P_i).(1-2P_j)}$$

$$P_i = \frac{M_i}{M}$$

con

$$D = \frac{1}{2}\left(1 + \sum_{i}^{N}\frac{P_i}{1-2P_i}\right)$$

Podemos seguir los cálculos necesarios, formación del espacio muestral y estimaciones para el esquema sin reposición y probabilidades desiguales al tamaño, utilizando una hoja de cálculo[5]. Para la formación del espacio muestral se ha considerado que las

[5] Los valores de la variable son Xi = (1, 2, 3) y los tamaños son Mi = (3, 4, 5). N = 3 y n = 2.

muestras que tienen las mismas unidades en distinto orden son idénticas.

Tabla 23.- Hoja de cálculo para un esquema de Horvitz-Thompson

HT[6]	1	2	3	4	5	6	7[7]	9	10	11	12[8]	13	14
	X1	X2	M1	M2	PI1	PI2	Xht	E(xht)	Var(Xht)	Plij	V(xhtm)	E(v)	V(V)
1.-	1	3	3	5	0,4	0,6667	7,00	0,4667	0,0667	0,0667	12,0000	0,8000	8,1718
2.-	1	4	3	7	0,4	0,9333	6,79	2,2619	0,4915	0,3333	0,3827	0,1276	0,0993
3.-	3	4	5	7	0,67	0,9333	8,79	5,2714	0,3704	0,6000	0,0017	0,0010	0,5155
Totales =							22,57	8,0000	0,9286	1,0000		0,9286	8,7866

La interpretación de las columnas es la siguiente:

Las cuatro primeras columnas tiene la misma interpretación que en el cuadro anterior, salvo que se han considerado iguales las muestras con los mismos elementos en distinto orden. Las columnas cinco y seis contienen los valores $p_i = nP_i$, necesarios para calcular en la columna siguiente el estimador del total. Las columnas nueve y diez son similares en su interpretación a las del cuadro anterior. La columna once calcula el factor π_{ij} . La columna doce calcula el estimador de la varianza para cada muestra.

En resumen, en este ejemplo tendremos que la varianza del estimador del total es 0,93 y la varianza del estimador de la varianza

[6] Muestras que es posible obtener con el procedimiento HT.
[7] Estimador del total para cada muestra con el método HT.
[8] Estimador de la varianza con el método HT.

es 8,79. También es posible obtener estos datos con la aplicación POSDEM. Recordamos que en el esquema con reposición y probabilidades proporcionales al tamaño la varianza del estimador del total era 1.14 y la varianza del estimador de la varianza es 2.72. por tanto y para los datos que se han utilizado como ejemplo el pasar de un método con reposición a otro sin reposición ha llevado aparejada una ganancia en precisión puesto que el error ha pasado de 1.14 a 0.93 y una pérdida en cuanto a estabilidad del error que pasa de 2.72 a 8.79.

Que la estimación del error tenga un elevado error de muestreo quiere decir que una vez obtenida una muestra, se pierde confianza en la estimación del error, debido a que los límites son más amplios. Por tanto si de obtener una muestra a obtener otra la estimación del error puede ser muy diferente, esto representa un serio inconveniente del método sin reposición y probabilidades proporcionales, que se suma a las dificultades de cálculo expuestas.

CAPÍTULO 10

MUESTREO CON PROBABILIDADES DESIGUALES (II)

10.1. MUESTREO CON REPOSICIÓN PARCIAL. ESQUEMA DE SÁNCHEZ-CRESPO Y GABEIRAS

Brewer (1983), señala que toda la literatura publicada sobre muestreo con probabilidades desiguales se ha referido prácticamente al muestreo con y sin reposición. El método con reposición parcial, SCG, proporciona, bajo un enfoque de superpoblación, una varianza que es siempre inferior a la de Hansen-Hurwitz y para $n=2$ coincide con la del método Horwtiz-Thompson. Es de fácil aplicación y el caso $n=2$ no supone una seria restricción ya que como dicen Rao y Bayless(1970) la selección de dos unidades por estrato es, con mucho, el caso más importante en las grandes encuestas. Brewer (1983) dice que $n=2$ unidades por estrato es un caso límite donde la máxima ventaja de la estratificación es consistente con la obtención de un estimador insesgado de la varianza. En análogo sentido se manifiesta Cochran (1977).

Cassel et al (1977:48) señalan que la estrategia HT puede ser rechazada como impracticable, incluso siendo óptima según ciertos criterios matemáticos.

En análogo sentido Brewer (1983: 109-110) dice que al no existir un procedimiento ideal para proceder en el muestreo sin

reposición y probabilidades desiguales, algunos pueden preferir el muestreo HH, por su facilidad en los procesos de selección y estimación, aunque se pierda completamente la reducción en varianza representada por el factor de corrección en poblaciones finitas.

En el muestreo con reposición, si se selecciona la unidad Ui con probabilidad Mi/M todos los elementos relativos a Ui vuelven a considerarse en la segunda selección. Si el muestreo es sin reposición todas las unidades Mi relativas a Ui se extraen de la población. En el esquema con reposición parcial, y considerando que en la primera extracción se ha seleccionado la unidad i, sólo una parte de los Mi elementos que hacen referencia a la unidad Ui se consideran en la segunda extracción. Concretamente, antes de la segunda selección se extraen de la urna b elementos, quedando por tanto en la urna Mi-b unidades que hacen referencia al elemento Ui. Donde b es igual al mayor entero de la expresión $(\frac{\text{Min. } M_i}{n-1})$

El estimador insesgado para el total es

$$\hat{X}_{SCG} = \sum_{i}^{n} \frac{X_i}{nP_i}$$

La varianza del estimador es:

$$V(\hat{X}_{SCG}) = \frac{M-nb}{M-b} * \frac{1}{n} \sum_{i}^{n} P_i (\frac{X_i}{P_i} - X)^2$$

Un estimador insesgado y no negativo para la varianza es:

$$\hat{V}(\hat{X}_{SCG}) = \frac{M-nb}{M} * \frac{1}{n(n-1)} * \sum_{i}^{n} (\frac{X_i}{P_i} - \hat{X}_{SCG})^2$$

Para n=2 se obtiene la expresión

$$\hat{V}(\hat{X}_{SCG}) = \frac{M-2b}{M} * \frac{1}{4} * (\frac{X_1}{P_1} - \frac{X_2}{P_2})^2$$

Se puede por tanto como hemos hecho en lo dos ejemplos anteriores formar el espacio muestral y las estimaciones para el esquema con reposición parcial y probabilidades proporcionales al tamaño. Así para la misma población, N=3 , n=2, Xi=1,3,4; Mi=3,5,7; tendremos:

Tabla 24.- Espacio muestral y las estimaciones para el esquema Sánchez-Crespo y Gabeiras.

SCG [1]	1	2	3	4	5[2]	6	7	8	9[3]	10	11
	X1	X2	M1	M2	Xsc	P(x)	E(xsc)	Var(Xsc)	V(xscm)	E(v)	V(V)
1.-	1	3	3	5	7,00	0,0833	0,5833	0,0833	2,4000	0,2000	0,1984
2.-	1	4	3	7	6,79	0,1167	0,7917	0,1720	1,9133	0,2232	0,1301
3.-	3	1	5	3	7,00	0,0833	0,5833	0,0833	2,4000	0,2000	0,1984
4.-	3	3	5	5	9,00	0,0556	0,5000	0,0556	0,0000	0,0000	0,0408
5.-	3	4	5	7	8,79	0,1944	1,7083	0,1200	0,0276	0,0054	0,1338
6.-	4	1	7	3	6,79	0,1167	0,7917	0,1720	1,9133	0,2232	0,1301
7.-	4	3	7	5	8,79	0,1944	1,7083	0,1200	0,0276	0,0054	0,1338
8.-	4	4	7	7	8,57	0,1556	1,3333	0,0508	0,0000	0,0000	0,1143
Totales =					62,71	1,0000	8,0000	0,8571		0,8571	1,0797

La interpretación de las primeras columnas es igual al caso anterior, variando el método de selección. En la columna cinco tenemos los valores estimados del total, que para la muestra (1,3) proporciona el valor:

$$\hat{X}_{sc} = \sum_i \frac{X_i}{n p_i} = 7$$

En la columna seis tenemos la probabilidad de obtención de cada muestra, que con el método de selección parcial se obtiene de la expresión

[1] Muestras que es posible seleccionar utilizando el método de SCG.
[2] Estimador del total obtenido para cada muestra con el procedimiento de SCG.
[3] Estimador de la varianza para cada muestra por el método de SCG.

$$P(\underline{x}) = \frac{M_i}{M} \frac{M_{j*}}{M - b}$$

en la que j es distinto de su valor para los casos siguientes

en la muestra $3,3$ $\frac{M_i}{M} \frac{M_i - b}{M - b}$

y en la muestra $4,4$ $\frac{M_i}{M} \frac{M_i - b}{M - b}$

En la columna siete tenemos la esperanza del estimador

$$E\,\hat{X}_{SC} = \hat{X}_{SC} \times P(\underline{x}) = 0.58 \quad \underline{\Sigma} = 8^{4}$$

En la columna ocho la varianza del estimador en base a la función de probabilidad

$$Var(\hat{X}_{SC}) = (\hat{X}_{SC} - E(\hat{X}_{SC}))^2 P(\underline{x}) = 0.08 \quad \underline{\Sigma} = 0.852$$

En la columna nueve el estimador para cada muestra

$$\hat{V}(\hat{X}_{SCM}) = \frac{M - bn}{M} \frac{1}{2} [\Sigma(\frac{X_i}{P_i} - \hat{X}_{SCG})^2] =$$

$$= \frac{M - b\,n}{M} \frac{1}{4} [\frac{X_1}{P_1} - \frac{X_2}{P_2}]^2 = 2.4$$

---

[4] Con la expresión _Sumatorio queremos indicar la suma extendida a toda la columna.

En las columnas diez y once la esperanza y la varianza de la varianza en base a la función de probabilidad

$$E(v) = \hat{V}(\hat{X}_{SCM}) \, x \, P(\underline{x}) = 0.2 \_ \Sigma = 0.857$$

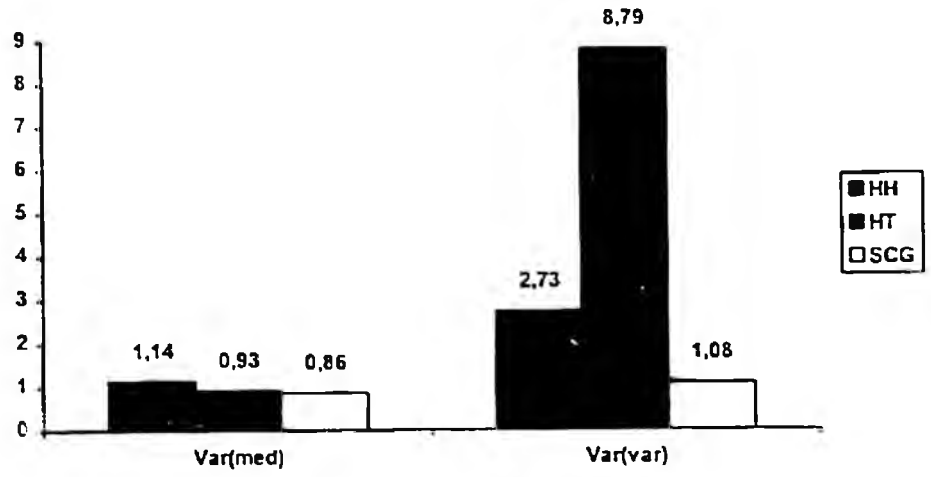
$$V(v) = (\hat{X}_{SCM} - E(\hat{X}_{SCM}))^2 \, P(\underline{x}) \_ 0.19 \_ \Sigma = 1.07$$

En resumen con el método de reposición parcial tendremos una varianza del estimador del total igual a 0,8571 y una varianza de la varianza igual a 1,0797

10.2. ANÁLISIS GRÁFICO DE COMPARACIÓN

Con los datos del ejemplo anterior es posible presentar de una forma ordenada los resultados en un gráfico. En este diagrama se puede observar como el método de Sánchez-Crespo y Gabeiras presenta, para esta población concreta, el menor error de muestreo, la menor varianza del total; y la mayor estabilidad de varianzas, esto es, la menor varianza de varianza.

Gráfico 11.- Varianza del estimador y estabilidad de varianzas para los esquemas HH HT y SCG.



Los valores de la variable son ..... Xi = (1, 3, 4)
Los tamaños son ..... Mi = (3, 5, 7)

Por tanto se había comprobado en el capítulo anterior que, para la estructura de población utilizada, el método sin reposición y probabilidades proporcionales al tamaño tenía un estimador más preciso, con menor error que el método con reposición, pero que a su vez tenía un error del error demasiado elevado, lo que suponía que de seleccionar una muestra a seleccionar otra los errores variaban excesivamente. Con el método con reposición parcial, y para estos datos, el estimador gana en precisión a los otros dos métodos, pero, además presenta una mayor estabilidad de varianzas.

No obstante estos resultados hacen referencia a una población concreta, vamos a seguir otro ejemplo para ver que cambios se producen sobre esta situación.

Vamos, aprovechando que está definida la estructura de la hoja de cálculo, a realizar otro ejemplo, esta vez con valores Xi=(1,2,3) y tamaños Mi=(3,4,5). Los resultados se pueden seguir en la tabla siguiente. Se podrían obtener los mismos resultados con la aplicación POSDEM, con la ventaja de no tener que limitar el tamaño de población.

Tabla 25.- Espacio muestral y las estimaciones para los esquemas HH, HT y SCG. Mediante una hoja de cálculo para los tres métodos y valores Xi=(1,2,3) con Mi=(3,4,5)

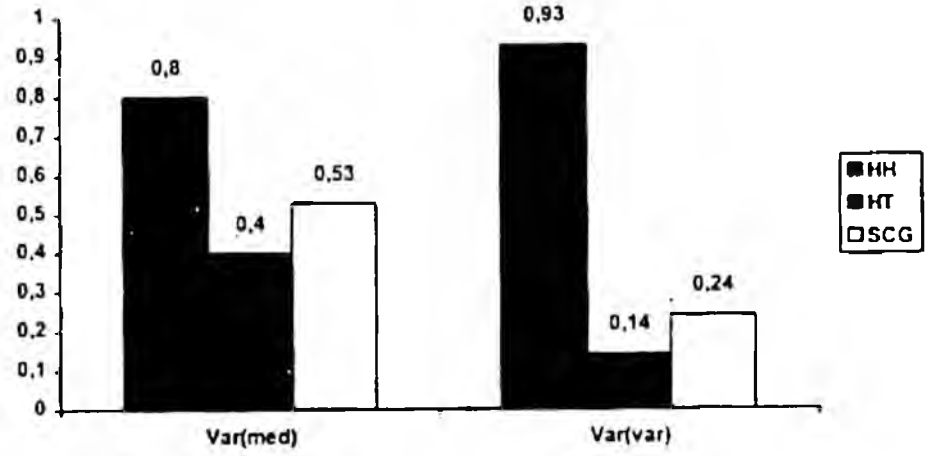
HH	1	2	3	4	5	6	7	8	9	10	11
	X1	X2	M1	M2	Xhh	P(x)	E(xhh)	Var(Xhh)	V(xhhm)	E(v)	V(V)
1.-	1	1	3	3	4,00	0,0625	0,2500	0 2500	0,0000	0,0000	0,0400
2.-	1	2	3	4	5,00	0,0833	0,4167	0,0833	1,0000	0,0833	0,0033
3.-	1	3	3	5	5,60	0,1042	0,5833	0,0167	2,5600	0,2667	0,3227
4.-	2	1	4	3	5,00	0,0833	0,4167	0,0833	1,0000	0,0833	0,0033
5.-	2	2	4	4	6,00	0,1111	0,6667	0,0000	0,0000	0,0000	0,0711
6.-	2	3	4	5	6,60	0,1389	0,9167	0,0500	0,3600	0,0500	0,0269
7.-	3	1	5	3	5,60	0,1042	0,5833	0,0167	2,5600	0,2667	0,3227
8.-	3	2	5	4	6,60	0,1389	0,9167	0,0500	0,3600	0,0500	0,0269
9.-	3	3	5	5	7,20	0,1736	1,2500	0,2500	0,0000	0,0000	0,1111
Totales =					51,60	1,0000	6,0000	0,8000		0,8000	0,9280

HT	1	2	3	4	5	6	7	9	10	11	12	13	14
	X1	X2	M1	M2	PI1	PI2	Xht	E(xht)	Var(Xht)	Plij	V(xhtm)	E(v)	V(V)
1.-	1	2	3	4	0,5	0,6667	5,00	0,8333	0,1667	0,1667	1,0000	0,1667	0,0600
2.-	1	3	3	5	0,5	0,8333	5,60	1,8667	0,0533	0,3333	0,6400	0,2133	0,0192
3.-	2	3	4	5	0,6667	0,8333	6,60	3,3000	0,1800	0,5000	0,0400	0,0200	0,0648
Totales =							17,20	6,0000	0,4000	1,0000		0,4000	0,1440

SCG	1	2	3	4	5	6	7	8	9	10	11
	X1	X2	M1	M2	Xsc	P(x)	E(xsc)	Var(Xsc)	V(xscm)	E(v)	V(V)
1.-	1	2	3	4	5,00	0,1111	0,5556	0,1111	0,5000	0,0556	0,0001
2.-	1	3	3	5	5,60	0,1389	0,7778	0,0222	1,2800	0,1778	0,0774
3.-	2	1	4	3	5,00	0,1111	0,5556	0,1111	0,5000	0,0556	0,0001
4.-	2	2	4	4	6,00	0,0370	0,2222	0,0000	0,0000	0,0000	0,0105
5.-	2	3	4	5	6,60	0,1852	1,2222	0,0667	0,1800	0,0333	0,0231
6.-	3	1	5	3	5,60	0,1389	0,7778	0,0222	1,2800	0,1778	0,0774
7.-	3	2	5	4	6,60	0,1852	1,2222	0,0667	0,1800	0,0333	0,0231
8.-	3	3	5	5	7,20	0,0926	0,6667	0,1333	0,0000	0,0000	0,0263
Totales =					47,60	1,0000	6,0000	0,5333		0,5333	0,2382

Para ver de una forma rápida estos resultados se ha realizado un diagrama de barras:

Gráfico 12.- Varianza del estimador y estabilidad de varianzas para los esquemas HH HT y SCG.



Los valores de la variable son ..... Xi = (1, 2, 3)
Los tamaños son ..... Mi= (3, 4, 5)

Aquí se aprecia como, para estos nuevos datos, el método de Horwitz-Thompson es más preciso y estable, que los otros dos métodos. No obstante hay que destacar como el método de Sánchez-Crespo-Gabeiras no presenta grandes variaciones en ninguno de los dos ejemplos. Para realizar un estudio de la estabilidad de procedimientos se ha incluido en un apartado siguiente un modelo de superpoblaciones.

10.3. PROBABILIDADES PROPORCIONALES Y LA APLICACIÓN POSDEM

Con esta aplicación se puede seguir los ejemplos anteriores, con la ventaja de poder utilizar otras poblaciones introducidas desde ficheros externos o bien desde una pantalla de la misma aplicación. El programa informático presenta similitudes con el utilizado para probabilidades iguales, no obstante es preciso resaltar que en este caso los estimadores hacen referencia al total por defecto, y por similitud con los manuales de muestreo consultados, si bien es posible obtener estimaciones para la media.

Con los tres métodos descritos en estos capítulos se pueden hacer prácticas con probabilidades proporcionales al valor, si bien con un carácter de comprobación, puesto que las estimaciones coinciden siempre con el valor del parámetro. No obstante dada su sencillez de programación, basta con cambiar la variable auxiliar por la variable de estudio, no supone un aumento de las líneas de programación.

A continuación se proporciona una salida impresa que contiene las estimaciones para cada muestra para el caso sin reposición y probabilidades proporcionales a los tamaños, que es el más complicado desde el punto de vista del cálculo.

Tabla 26.- Salida impresa de POSDEM que contiene las estimaciones para cada muestra para el caso sin reposición y probabilidades proporcionales al tamaño.[5]

Filas[6] \ Columnas [7]	TOTAL	VAR	DES	Li	Ls	Cvs(%)
ESPERANZA	5.89	0.47	0.68	4.52	7.26	12
VARIANZA...	0.45	0.16	0.12	1.88	0.01	52
DESVIACION	0.67	0.40	0.35	1.37	0.09	7
CO.VAR .(%)	11	86	60	29	1	67
LIM.SU....	7.23	1.27	1.29	7.46	7.25	25
LIM.IF....	4.55	-0.34	-0.12	1.98	6.88	-4

Filas[8] \ Columnas [9]	TOTAL	VAR	DES	Li	Ls	Cvs(%)
Muestra(1)	5.00	1.00	1.00	3.00	7.00	20
Muestra(2)	5.60	0.64	0.80	4.00	7.20	14
Muestra(3)	6.60	0.04	0.20	6.20	7.00	3
Muestra(4)	6.60	0.04	0.20	6.20	7.00	3
Muestra(5)	6.60	0.04	0.20	6.20	7.00	3
Muestra(6)	5.60	0.64	0.80	4.00	7.20	14
Muestra(7)	6.60	0.04	0.20	6.20	7.00	3
Muestra(8)	6.60	0.04	0.20	6.20	7.00	3

[5] Los valores de la variable son Xi = (1, 2, 3) y los tamaños son Mi= (3, 4, 5)
[6] Esperanza del estimador del total calculada sobre un representación del espacio muestral. En las siguientes filas se calculan la varianza y otros parámetros deducidos.
[7] Estimador del total, de la varianza y de otros parámetros.
[8] En filas se representan los resultados obtenidos para las primeras ocho muestras generadas.
[9] En columnas los estimadores del total, la varianza y otros deducidos.

10.4. VARIANZAS ESPERADAS CON UN MODELO DE SUPERPOBLACIÓN

Vamos a aplicar el modelo de superpoblación

$X_i = B M_i + e_i$ con $i = 1, 2, 3, ..., N$ e_i es una variable aleatoria y E^* es el valor esperado sobre todas las posibles poblaciones finitas que hipotéticamente pueden deducirse del modelo, condicionadas a un conjunto fijo de M_i valores.

$E^* (e_i / M_i) = 0$ $E^*_{i,j} (e_i e_j / M_i M_j) = 0$ $E^*(e_i^2 (M_i)) = a M_i^g$

con

$a > 0$, $1 \leq g \leq 2$

La reducción esperada de varianza para el esquema de SCG sobre el esquema de HH es[10]:

$$R_e = 1 - \frac{E^* V(\hat{X}_{SCG})}{E^* V(\hat{X}_{HH})} = \frac{b(n-1)}{M-b} = \frac{3(2-1)}{12-3} = \frac{1}{3}$$

Y para el método HT tenemos

[10] Los valores de la variable son Xi = (1, 2, 3) y los tamaños son Mi = (3, 4, 5)

$$R_1 = 1 - \frac{E^*V(\hat{X}_{HT})}{E^*V(\hat{X}_{HH})} = \frac{n-1}{N-1} = \frac{1}{2}$$

Ei ratio de reducción en varianza esperada con SCG a la reducción con HT, ambas en relación con HH es:

$$\frac{1 - \frac{E^*V(\hat{X}_{SCG})}{E^*V(\hat{X}_{HH})}}{1 - \frac{E^*V(\hat{X}_{HT})}{E^*V(\hat{X}_{HH})}} = \frac{1 - \frac{b(n-1)}{M-b}}{1 - \frac{\frac{aM}{n}}{\frac{aM}{n}} \frac{N-n}{N-1}} = \frac{\frac{b(n-1)}{M-b}}{\frac{n-1}{N-1}} = \frac{\frac{b(n-1)}{N\overline{M}-b}}{\frac{n-1}{N-1}} = \frac{b(N-1)}{N\overline{M}-b}$$

y si consideramos la definición de b entonces tendremos

$$\rho = \frac{N-1}{N\frac{\overline{M}}{M_o}(n-1)-1} \leq \frac{N-1}{N(n-1)-1}$$

donde $M_o = \min. M_i$ y $\frac{\overline{M}}{M_o} \leq 1$

En Sánchez-Crespo (1994:152) puede estudiarse la máxima reducción en varianza esperada para distintos valores de n y N.

Tabla 27.- Reducción en varianza esperada para distintos valores de tamaño de población y de muestra

n = 2	N = 3	ρ = 1.0
	N = 20	ρ = 1.0
n = 10	N = 20	ρ = 0.49

Así la máxima reducción en varianza esperada se produce para n=2 en el esquema SCG, en relación con el HT.

La proporción del ahorro potencial en varianza esperada, se puede obtener por el siguiente indicador propuesto por Brewer

$$\rho = \frac{E^*V(\hat{X}_{HH}) - E^*V(\hat{X}_{SCG})}{E^*V(\hat{X}_{HH}) - E^*V(\hat{X}_{HT})} = \frac{b(N-1)}{M-b} \qquad 1 \geq \rho > 0$$

En el ejemplo $\rho = \frac{2}{3}$.

Cuando la esperanza de la varianza del estimador de SCG es igual a la correspondiente expresión de HT entonces ρ tomará el valor uno. El valor cero es inaccesible por ser siempre la esperanza de la varianza de SCG menor que la correspondiente expresión de HH. Así cuanto más próximo a uno esté el indicador el método SCG estará más próximo al HT. Por último si ρ>0.5 la varianza esperada de SCG estará más próxima al HT que al HH.

10.5. ESTABILIDAD DEL ESTIMADOR DE LA VARIANZA ESPERADA PARA LOS TRES PROCEDIMIENTOS CONSIDERANDO EL INDICADOR DE RAO Y BAYLESS

El indicador tiene la siguiente formulación:

$$I^2_{RB}\, E\left(\hat{V}(\hat{X})\right) = \frac{E*E\hat{V}^2(\hat{X}) - \left[E*E\hat{V}(\hat{X})\right]}{\left(E*E\hat{V}(\hat{X})\right)^2} = \frac{E*E\hat{V}^2(\hat{X})}{\left(E*E\hat{V}(\hat{X})\right)^2} - 1$$

Para $g=2$ (caso más desfavorable para SCG) y $n=2$ los momentos de 2° orden son:

$$E*E\hat{V}^2\left(\hat{X}_{HT/B}\right) = 3a^2M^4 \cdot \frac{1}{2} \cdot \sum_{i}^{N}\left(4P_iP_i - \pi_{ij}\right)^2$$

$$E*E\hat{V}^2\left(\hat{X}_{HH}\right) = 3a^2M^4 \cdot \frac{1}{2} \cdot \sum_{i<j}^{N} P_iP_j$$

$$E*E\hat{V}^2\left(\hat{X}_{SCG}\right) = \left(\frac{M-2b}{M}\right)^2 E*E\hat{V}^2\left(\hat{X}_{HH}\right) = \left(\frac{M-2b}{M}\right)^2 \cdot \frac{3a^2M^4}{2} \sum_{i<j}^{N} P_iP_j$$

y para los denominadores de $I_{RB}^2 E\left(\hat{V}(\hat{X})\right)$

$$\left(E * E\left(V\left(\hat{X}_{HT/B}\right)\right)\right)^2 = a^2 M^4 \frac{1}{4}\left(1 - 2\sum_i^N P_i^2\right)^2$$

$$\left(E * E\left(V\left(\hat{X}_{HH}\right)\right)\right)^2 = a^2 M^4 \frac{1}{4}\left(1 - \sum_i^N P_i^2\right)^2$$

$$\left(E * E\left(V\left(\hat{X}_{SCG}\right)\right)\right)^2 = \left(\frac{M-2b}{M-b}\right)^2 \cdot \frac{a^2 M^4}{4}\left(1 - \sum_i^N P_i^2\right)^2$$

y sustituyendo en $I_{RB}^2 E\left(\hat{V}(\hat{X})\right.$ tenemos:

$$I_{RB}^2\left(\hat{V}\left(\hat{X}_{HH}\right)\right) = \frac{6\sum_{i<j}^N P_i P_j}{\left(1 - \sum_i^N P_i^2\right)^2} - 1$$

$$I_{RB}^2\left(\hat{V}\left(\hat{X}_{SCG}\right)\right) = \frac{\left(\frac{M-b}{M}\right)^2 \cdot 6\sum_{i<j}^N P_i P_j}{\left(1 - \sum_i^N P_i^2\right)^2} - 1$$

SCG es siempre más estable que HH.

$$I_{RB}^2 E\left(\hat{V}\left(\hat{X}_{HT/B}\right)\right) = \frac{3\sum_{i<j}^N \sum \left(4P_iP_j - \pi_{ij}\right)^2 / \pi_{ij}}{\left(1 - 2\sum_i^N P_i^2\right)^2} -$$

Ejemplo de aplicación: estabilidad de varianzas, indicadores de Rao y Bayless

$$I^2_{RB}\left(\hat{V}(\hat{X}_{HH})\right)=\frac{6\sum_{i<j}^{N} P_iP_j}{\left(1-\sum_{i}^{N} P_i^2\right)^2}-1 \qquad P_i\left(\frac{3}{12},\frac{4}{12},\frac{5}{12}\right)$$

$$6\sum_{i<j}^{N} P_iP_j=\left(\frac{3}{12}\cdot\frac{4}{12}+\frac{3}{12}\cdot\frac{5}{12}+\frac{4}{12}\cdot\frac{5}{12}\right)6=\frac{282}{144}=1{,}9583$$

$$\left(1-\sum_{i}^{N} P_i^2\right)^2=\left(1-\frac{50}{144}\right)^2=0.426 \qquad I^2_{RB}\left(\hat{V}\left(\hat{X}_{HH}\right)\right)=3.60$$

$$I^2_{RB}\left(\hat{V}(\hat{X}_{SCG})\right)=\left(\frac{M-b}{M}\right)^2\cdot\frac{6\sum_{i<j}^{N} P_iP_j}{\left(1-\sum_{i}^{N} P_i^2\right)^2}-1=\frac{81}{144}\cdot\frac{1.9583}{0.426}-1=1.59$$

$$\left(\frac{M-b}{M}\right)^2=\left(\frac{9}{12}\right)^2$$

$$I^2_{RB}\left(\hat{V}(\hat{X}_{HT})\right)=\frac{3\sum_{i<j}^{N}\left(4P_iP_j-\pi_{ij}\right)^2/\pi_{ij}}{\left(1-2\sum_{i}^{N} P_i^2\right)^2}-1$$

$$\left.\begin{array}{ll} 4P_1P_2 = \dfrac{48}{144} = 0.33 & \dfrac{(0.33-0.16)^2}{\frac{1}{6}} = 0.1734 \\ 4P_1P_2 = \dfrac{60}{144} = 0.42 & \dfrac{(0.42-0.32)^2}{\frac{2}{6}} = 0.0300 \\ 4P_2P_3 = \dfrac{80}{144} = 0.56 & \dfrac{(0.56-0.5)^2}{\frac{3}{6}} = 0.0072 \end{array}\right\} \quad 0.2106 x 3 = 0.6318$$

$$\left(1-2\sum_{i}^{N} P_i^2\right)^2 = \left(1-\frac{100}{144}\right)^2 = \left(\frac{44}{144}\right)^2 = 0.0934$$

$$I_{RB}^2\left(\hat{V}(\hat{X}_{HT})\right) = \frac{0.6318}{0.0934} = 6.76$$

Por tanto

$$I_{RB}^2(\hat{V}(\hat{X}_{SCG})) < I_{RB}^2(\hat{V}(\hat{X}_{HH})) < I_{RB}^2(\hat{V}(\hat{X}_{HT}))$$

Según estos indicadores el esquema SCG es más estable. Rao y Bayles han propuesto este indicador porque el valor esperado del cuadrado del coeficiente de variación

$$E^*(CV^2(\hat{V}(\hat{X}))) = E^*\frac{V(\hat{V}(\hat{X}))}{E^*V(\hat{X})^2}$$

es un ratio de un cociente con dos variables aleatorias y no es tan sencillo de obtener.

10.6. OTRAS PROPIEDADES DEL ESQUEMA SCG

a) Propiedad del estimador de la razón: la expresión para la varianza del estimador del total

$$V(\hat{X}_{SCG}) = \frac{M-nb}{M-b}\,\frac{1}{n}\sum_{i}^{N}[(\frac{X_i}{P_i})-X]^2 P_i$$

es igual a cero, cuando M_i es exactamente proporcional al valor desconocido X_i.

b) La propiedad de rotabilidad[11]: la probabilidad en la primera selección para la unidad u_i es

$$P(u_i;1^{a}selc.) = P_i$$

y en la segunda selección es:

$$P(u_i;2^{a}selc.) = \frac{M_i}{M}\,\frac{M_i-b}{M-b} + \frac{M-M_i}{M}\,\frac{M_i}{M-b} = \frac{M_i^2-b\,M_i + M\,M_i - M_i^2}{M(M-b)} = \frac{M_i}{M} = P$$

Por tanto la probabilidad incondicional de seleccionar u_i en primera o en segunda selección es igual a P_i, y la probabilidad de seleccionar u_i en una muestra de tamaño dos es: $P(u_i \in \text{muestra}) = P_i + P_i = 2P_i$

[11] Esta propiedad es especialmente importante en encuestas continuas con rotación. Brewer (1983:68-71).

10.7. MÉTODO DE SELECCIÓN DE BREWER PARA EL ESQUEMA SIN REPOSICIÓN

Vamos a considerar el caso n=2 y $P_i < \frac{1}{2}$ siguiendo a Cochran (1977:261-263).

En la primera selección con probabilidad proporcional a $\frac{P_i(1 - P_i)}{(1 - 2P_i)D}$ donde D es el divisor necesario en orden a que la expresión sea una probabilidad real

$$D = \sum_i^N \frac{P_i(1 - P_i)}{1 - 2P_i} = \frac{1}{2} \sum_i^N \frac{P_i(2 - 2P_i)}{1 - 2P_i} = \frac{1}{2} \sum_i^N \frac{P_i(1 + 1 - 2P_i)}{1 - 2P_i} =$$

$$= \frac{1}{2} \sum_i^N \frac{P_i}{1 - 2P_i} + \frac{1}{2} \sum_i^N \frac{P_i(1 - 2P_i)}{1 - 2P_i} = \frac{1}{2} \sum_i^N \frac{P_i}{1 - 2P_i} + \frac{1}{2} \sum_i^N P_i$$

$$= \frac{1}{2} \left(\sum_i^N \frac{P_i}{1 - 2P_i} + 1 \right)$$

En la segunda selección

con probabilidad $\frac{P_i}{1} - P_i$ donde u_i es la unidad seleccionada

Probabilidad de seleccionar la unidad u_i será $2P_i$

p_i = **probabilidad de que la unidad** u_i **sea seleccionada con** $n=2$

$$= \frac{P_i(1-P_i)}{(1-2P_i)D} + \frac{1}{D}\sum_{j\neq i}^{N}\frac{P_i(1-P_j)}{1-2P_j}\frac{P_i}{1-P_i} =$$

$$= \frac{P_i}{D}\left[\frac{1-P_i}{1-2P_i} + \frac{1}{D}\sum_{j\neq i}^{N}\frac{P_j(1-P_j)}{1-2P_j}\frac{1}{(1-P_j)}\right] =$$

$$= \frac{P_i}{D}\left[\frac{1-P_i-P_i+P_i}{1-2P_i} + \sum_{j\neq i}^{N}\frac{P_j}{1-2P_j}\right] = \frac{P_i}{D}\left[\frac{1-2P_i+P_i}{1-2P_i} + \sum_{j\neq i}^{N}\frac{P_j}{1-2P_j}\right] =$$

$$= \frac{P_i}{D}\left[1+\frac{P_i}{1-2P_i} + \sum_{j\neq i}^{N}\frac{P_j}{1-2P_j}\right] = \frac{P_i}{\frac{1}{2}\left[\sum_{i}^{N}\frac{P_i}{1-2P_i}+1\right]}\left[1+\sum_{i}^{N}\frac{P_i}{1-2P_i}\right] = 2P_i$$

Probabilidad de que u_i y u_j estén ambas en la muestra

$$\pi_{ij} = \frac{2P_iP_j}{D}\frac{(1-P_i-P_j)}{(1-2P_i)(1-2P_j)} \qquad \textbf{con}$$

$$\sum_{i}^{N}\sum_{j>i}\pi_{ij} = \frac{n(n-1)}{2} = 1$$

En el ejemplo, el valor de D = 2.5. Donde los cálculos necesarios para obtener el valor de D son con los datos del ejemplo:

$$D = \frac{1}{2}\left(\sum_{i}^{N}\frac{P_i}{1-2P_i}+1\right) = \frac{1}{2}\left(\frac{48}{12}+1\right) = \frac{1}{2}\,\frac{60}{12} = \frac{5}{2} = 2.5$$

$$\pi_{12} = \frac{2P_1P_2}{2.5}\frac{(1-P_1-P_2)}{(1-2P_1)(1-2P_2)} = \frac{1}{6}$$

$$\pi_{13} = \frac{2 P_1 P_3}{2.5} \frac{(1 - P_1 - P_3)}{(1 - 2 P_1)(1 - 2 P_3)} = \frac{2}{6}$$

$$\pi_{23} = \frac{2 P_2 P_3}{2.5} \frac{(1 - P_2 - P_3)}{(1 - 2 P_2)(1 - 2 P_3)} = \frac{3}{6}$$

10.8. EXTENSIÓN UTILIZADA EN POSDEM AL CASO DE MUESTRAS DE TAMAÑO MAYOR QUE DOS

Vamos a enfocar este apartado con un ejemplo. Los datos del problema son: utilizando una población de tamaño 1600, obtener muestras de tamaño 40 con los diferentes esquemas de probabilidades desiguales estudiados. Hemos visto que por necesidades de la teoría el muestreo con probabilidades desiguales es más eficiente cuando n=2, para solventar esta dificultad hemos recurrido en este caso al artificio de considerar la población dividida en 20 grupos, estratos, de 80 unidades cada uno. En cada grupo seleccionaremos una muestra independiente de tamaño igual 2. De esta forma por agregación tendremos un total de 40 unidades muestrales.

Las fórmulas de aplicación en este caso son:

$$\bar{\bar{x}}_{st} = \sum_{h}^{L} wh \times \bar{\bar{x}}_h = \sum_{h}^{L} \frac{Nh}{N} \bar{\bar{x}}_h$$

$$V(\bar{\bar{x}}_{st}) = \sum_{h}^{L} w_h^2 \; V(\bar{\bar{x}}_h) = \sum_{h}^{L} \left(\frac{Nh}{N}\right)^2 \cdot V(\bar{\bar{x}}_h)$$

Aplicadas al estimador del total , tenemos:

$$\hat{x}_m = N \cdot \hat{\bar{x}}_m = N \cdot \sum_{h}^{L} \frac{Nh}{N} \cdot \hat{\bar{x}}_h = \sum_{h}^{L} Nh \, \hat{\bar{x}}_h = \sum_{h}^{L} \hat{x}_h$$

$$V(\hat{x}_m) = N^2 V(\hat{\bar{x}}_m) = \sum_{h}^{L} Nh^2 \cdot V(\hat{\bar{x}}_h) = \sum_{h}^{L} V(\hat{x}_h)$$

Así, obtenidos los valores de $\hat{\bar{x}}_h$ y de $V(\hat{x}_h)$ con tamaños de muestra en cada estrato igual a dos unidades y las fórmulas de probabilidades desiguales de los capítulos anteriores, podremos obtener las estimaciones para el total y su varianza correspondientes a tamaños de muestra superiores a dos.

CAPÍTULO 11

DATOS

11.1. INTRODUCCIÓN

El marco en las encuestas por muestreo se puede definir como el conjunto de información disponible para realizar los procesos de selección y estimación. Permitirá determinar el alcance y los parámetros básicos de la investigación. En el caso de España, para las encuestas oficiales por muestreo se utilizan básicamente dos variantes según se trata de encuestas demográficas o de encuestas económicas.

1.1.- Para las «encuestas demográficas» se utiliza el marco compuesto por las aproximadamente 30.000 secciones censales. Estas secciones, dentro del diseño de la Encuesta General de Población[1], se agrupan en estratos y se obtiene una muestra de 3000 secciones como unidades de primera etapa. Sobre cada una de estas secciones se realiza un muestreo sistemático que permite disponer de 20 viviendas por sección, que es el trabajo semanal de un entrevistador. La selección de estas últimas 20 viviendas se lleva a cabo utilizando como marco el cuaderno de tabulación manual del agente censal. En el cuaderno de tabulación figura, además de otros datos, un número de identificación y el número de personas que habitan en el hogar y la dirección de la vivienda. Este cuaderno se realiza con ocasión de los censos de población y cada renovación quinquenal. Para evitar un desfase

[1] Las principales características de la Encuesta General de Población puede verse en García España (1974).

importante debido al paso del tiempo este marco se actualiza periódicamente para cada sección que pertenece a la muestra. En este trabajo vamos a utilizar el marco formado por las 1600 secciones de la comunidad del País Vasco con información procedente del censo de población referido a 1991. Existe una base de datos informatizada , basada en tecnología de Cd-rom , y publicada por el INE con el titulo "CERCA +100: Cifras de áreas pequeñas".

1.2.- Para las «encuestas económicas» se dispone del Directorio Centralizado de empresas, DIRCE. En este directorio figuran, entre otros datos, los relativos a actividad y número de trabajadores. Normalmente en las encuestas económicas se investigan exhaustivamente todas las empresas con más de veinte trabajadores, el proceso de muestreo se limita a las empresas con menos de veinte trabajadores. No obstante en el capítulo donde se analiza la aplicación de POSDEM a los datos de la encuesta industrial se ha considerado como marco de muestreo todas las empresas de un determinado estrato, tratándose de forma diferenciada la detección de unidades que deben ser tratadas como autorepresentadas, esto es de aquellas empresas que tienen probabilidad uno de pertenecer a la muestra.

Con la aplicación POSDEM se pueden resolver problemas para optimizar la selección y estimación de muestras con datos que proceden de ficheros externos. También se puede trabajar con datos introducidos por pantalla y existe también la posibilidad de definir datos automáticamente. Con esta aplicación resolveremos problemas con datos demográficos y económicos reales. También se realizaran prácticas

con datos generados aleatoriamente de poblaciones que se distribuyen normal o binomialmente. Por último, y más importante, es posible resolver problemas utilizando modelos complejos de superpoblación.

11.2. DATOS PROCEDENTES DE POBLACIONES NATURALES

POSDEM permite utilizar poblaciones reales o introducidas por el usuario mediante ficheros externos o directamente utilizando una pantalla de la aplicación.

Desde las opciones del programa podemos cargar en la memoria los datos correspondientes al marco de una encuesta. Utilizamos el concepto de marco en un sentido restringido, no como toda la información disponible para realizar una encuesta, sino como el conjunto de información disponible para cada unidad de la población. Se materializa en un listado con información sobre la identificación de cada unidad y los valores de las variables que vamos a utilizar en el análisis. Concretamente distinguiremos entre variable de estudio y variable auxiliar. La variable de estudio será aquella para la que se calculen los estimadores y sus errores de muestreo asociados y la variable auxiliar será aquella que se utilice en el proceso de selección y de estimación a la hora de asignar probabilidades variables para cada unidad de la población.

La población podrá estar ordenada o no según diferentes criterios: en base a la variable auxiliar, en base a la variable de estudio, o en base al orden de presentación. Para realizar comprobaciones de los

cálculos se pueden utilizar diferentes conjuntos de datos obtenidos de manuales de muestreo o de artículos publicados. Estos ficheros se han incorporado para ejemplo en la aplicación POSDEM en su apartado de datos. Entre los ficheros que se han utilizado para analizar los resultados de la aplicación POSDEM están:

1.- «Franjas de bosque y volumen de madera». Variable de estudio: "el volumen de madera". Tamaño de población: 176 "franjas de bosque". Variable auxiliar: " el ancho de franja". En esta población se espera que el volumen de madera esté relacionado con el ancho de franja. Ordenar la población según el ancho de franja será similar a ordenar conforme al volumen de madera. Los datos se encuentran en (Murthy,1967:131) y los resultados en la página149.

Se puede observar en estos resultados que, efectivamente, el comportamiento de la varianza del muestreo sistemático es irregular, al contrario de lo que ocurre con el muestreo sin reposición. Se puede observar también como el ordenar la población ha supuesto una ayuda considerable para reducir la varianza.

2.- Aldeas y censos de 1951 y 1961. EL tamaño de la población es de 128 aldeas. Las variables del censo de 1951 son: "área demográfica en millas cuadradas", "área cultivada en acres",

"núm. de personas". Las variables del censo 1961: son "número de personas", "número de cultivadores", "trabajadores en industrias familiares" y "número de hogares". Las variables de estudio son: "área cultivada" y "número de personas del censo de 1961". La población se considera con tres criterios diferentes: en primer lugar como en el marco, después en orden creciente del "área geográfica", y por último, según el "número de personas en 1951" en orden creciente. Los datos figuran en (Murthy,1967:128), y los resultados en la página 151.

Se puede observar que la ordenación de las unidades poblacionales más eficiente es el del "área geográfica" para estimar el "área cultivada", mientras que las ordenaciones según "número de personas del año 1951" es más eficiente que los otros dos tipos de ordenación para estimar la población de 1961. Otro resultado importante consiste en observar como, cuando la ordenación es la del marco, el muestreo sistemático es menos eficaz que el muestreo sin reposición para tamaños de muestra pequeños.

3.- «Fábricas, número de trabajadores, capital y producción». Tamaño de población 80 fábricas. página 228 del mismo autor.

4.- «Censo de población de 1991 por secciones censales en la Comunidad Autónoma de Cantabria». Las variables consideradas han sido las relativas a: "número de personas censadas", el número de personas en relación con la actividad: "número de activos", "de parados", "de ocupados", y parados por edad "menores de 16 años" y "mayores de 16 años". El número de secciones es de 402 correspondientes al total de la Comunidad Autónoma de Cantabria. Se ha utilizado también una población definida de igual forma para la Comunidad Autónoma del País Vasco con un tamaño de 1600 secciones censales, clasificadas según cada una de las tres provincias que la componen.

5.- «Hogares en una determinada sección censal». Variable considerada: "número de personas en el hogar". Tamaño de población: 400 hogares.

6.- «Empresas Industriales por número de trabajadores». Tamaño de población: 160 empresas. Variable de estudio "número de trabajadores".

7.- «Secciones Censales clasificadas por número de hogares en la sección». Tamaño de población: 20 secciones. Variable de estudio "número de hogares".

8.- «Secciones Censales clasificadas por número de personas en la sección». Tamaño de población: 16 secciones. Variable de estudio "número de parados censados en 1991" y variable auxiliar "población de la sección".

11.3. DATOS INTRODUCIDOS POR PROGRAMA. MODELOS DE SUPERPOBLACIÓN

Para introducir datos por pantalla se debe utilizar la opción correspondiente al editor de datos, que permite definir e introducir los datos del problema en la aplicación, bien uno a uno por pantalla, o bien de forma automática. Es posible definir de forma automática variables con fórmulas de grado de complejidad variable. Se han introducido en este apartado del programa diferentes posibilidades. No obstante en el caso de que se desee realizar un experimento con datos que sigan una determinada distribución de probabilidad es preferible generar estos datos con algunos de los programas de análisis estadístico disponibles que cumpla ese cometido.

Una vez que la aplicación trabaja con un cierto conjunto de datos que representan una población natural, es posible modelar la misma utilizando el modulo que hemos denominado "simulador de estructuras poblacionales". Con este modulo es posible determinar ajustes por polinomios ortogonales para diferentes tramos en la población y con distintos términos de perturbación aleatoria. El término de perturbación puede distribuirse bien de forma normal con media y varianza dadas o incluyendo un término de heterocedasticidad que hemos observado frecuente en las poblaciones analizadas.

11.4. PRÁCTICAS RESUELTAS CON DATOS CUANTITATIVOS: ENTRADA POR PANTALLA

1.- Utilizando un esquema de selección sistemática con intervalo variable y un esquema aleatorio sin reposición y fijados el tamaño de población en 24 unidades y el tamaño de muestra en 6, y el número de reiteraciones en 4 y en 40 para cada método.

Las unidades están identificadas por un número correlativo:

Ui = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

y, donde los valores de la variable para cada unidad es un información conocida, y que toma respectivamente los valores

x_i = 1 3 3 4 4 4 5 6 7 9 10 11 12 12 15 16 20 21 21 22 23 23 23 23

Se pide, paso a paso, con la aplicación POSDEM:

a.-) Los límites de confianza para el estimador al 95%, para muestreo sistemático con intervalo variable.

b.-) Establecer para el caso de muestreo aleatorio sin reposición los tamaños de muestra necesarios para obtener errores de muestreo entre el 1 % y el 20 %.

c.-) El error de muestreo para muestreo aleatorio sin reposición y tamaño de muestra igual a seis.

Solución:

- El primer paso, es introducir los datos mediante el editor de la aplicación.
- Después se determina el tamaño de población con el que se va a trabajar (24), el tamaño de muestra (6) y el número de reiteraciones (4). En este caso el tamaño de muestra es un dato del problema y el número de reiteraciones al utilizar un método sistemático se fija de forma obligada en 4. Cuando se utilice muestreo aleatorio sin reposición se fija el número de reiteraciones libremente sin más restricciones que la memoria del ordenador con el que estemos trabajando. POSDEM almacena los nuevos valores y permite presentar un gráfico con los valores introducidos.
- Después decidiremos el método elegido, en este caso muestreo sistemático con intervalo variable.
- Ahora, ya es posible con la opción listados obtener los datos que pide el problema:

a.-) Los límites de confianza del 95 para el estimador sistemático con intervalo variable son (11.8 y 13.0).

b.-) Determinaremos para el caso de muestreo aleatorio sin reposición los tamaños de muestra necesarios para obtener errores de muestreo para el estimador entre el 1 % y el 20 %.

La salida del programa al trabajar con los datos y las especificaciones dadas es la siguiente:

POSDEM

Comparación del muestreo sistemático con intervalo variable con el muestreo aleatorio sin reposición y probabilidades iguales.

Datos del diseño

Población marco utilizada : E1-24.dbf[2]
Variable de estudio : Definida arbitrariamente por el usuario
Variable auxiliar : Ninguna
Media poblacional = 12.4167 Desviación típica = 7.7186
Tamaño de poblacion = 24
Tamaño de muestra = 6
Fracción de muestreo = 25%
Error de muestreo CV% (Basado en las reiteraciones) = 2.4%

El tamaño necesario en el muestreo sin reposición hubiese sido para distintos coeficientes de variación ...

El tamaño de la muestra para la media +- 20 % = 7
El tamaño de la muestra para la media +- 15 % = 10
El tamaño de la muestra para la media +- 10 % = 15
El tamaño de la muestra para la media +- 7 % = 18
El tamaño de la muestra para la media +- 5 % = 21
El tamaño de la muestra para la media +- 3 % = 23
El tamaño de la muestra para la media +- 2 % = 23
El tamaño de la muestra para la media +- 1 % = 24

[2] Nombre del fichero que contiene los datos de la población natural.

Resultados para un error del 2.4 % el tamaño de muestra es = 23
La ganancia en tamaño de muestra, ((nsr-nm)/nsr), es = 73.9130%

La varianza de la media para el caso de sistemático con intervalo variable que para esta población es igual a ..... 0.0903 frente a la varianza de la media para el caso de muestreo estratificado con una unidad muestral en cada estrato es igual a 0.2153

Por tanto la ganancia ó pérdida en precisión del sistemático con intervalo variable según la diferencia sea positiva o negativa frente ai muestreo estratificado con $n_h = 1$ es del 58%

c.-) El error de muestreo para muestreo aleatorio sin reposición y tamaño de muestra igual a seis se sitúa en torno a 21.8% . Se observa que para el método sistemático con intervalo variable el error de muestreo es 1.9% .

2) Si transcurrido un periodo de tiempo determinado se obtiene una nueva muestra utilizando muestreo sistemático con intervalo variable y proporciona una estimación que presenta una disminución del 1% establecer si se trata de una disminución significativa o si se trata de una variación debida al proceso de muestreo.

Solución:

- Con los datos de este supuesto en lugar de 12.4 se habría obtenido una estimación igual a 11,1 si el error de muestreo fuese similar, se obtiene una desviación típica del estimador igual 0,24, y el intervalo de confianza tiene los límites 10,6 y 11,58. Por tanto como los límites anteriores eran 11.8 y 13.0, la variación que se ha producido, una

caída del 1% es significativa y no se ha producido como consecuencia de la variabilidad del proceso de muestreo.

3) Con los datos de la primera práctica desordenados, calcular para el muestreo sistemático con intervalo variable los límites de confianza para el 95% para el estimador de la media.

Ui = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

y, donde los valores de la variable para cada unidad es un información conocida, y que toma los valores

x_i = 7 21 5 6 22 23 9 10 11 12 12 1 3 4 15 3 23 23 4 4 16 20 21 23

Solución:

- Al utilizar un método sistemático sobre un fichero con datos en orden aleatorio el aplicar un método sistemático es equivalente a utilizar un método aleatorio simple sin reposición.

a.-) Límites (6.2 y 18.3)

11.5. PRÁCTICAS CON DATOS REALES CUALITATIVOS: ENTRADA POR PANTALLA

1) Con los siguientes datos

Ui =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
x_i=	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

donde el valor de la variable que corresponde a la unidad Ui toma los valores cero o uno en función de la ocurrencia o no de una determinada circunstancia.

Se pide:

a.-) Dado un tamaño de muestra igual a cuatro, calcular considerando todas las muestra sistemáticas con intervalo variable posibles, el valor esperado del estimador de la proporción y su error de muestreo

b.-) Para un conjunto de 40 muestras aleatorias simples sin reposición se quiere determinar el valor esperado de la varianza y la varianza de la varianza.

Solución:

- La primera pregunta se deduce del listado de cálculos para cada muestra. Y los resultados pedidos son 0.375 y 0.015625
- Y la segunda cuestión del listado de esperanzas y varianzas correspondiente: la media de las varianzas para el conjunto de las 40 reiteraciones es 0.0319 y su varianza, varianza del estimador de la varianza es a su vez igual a 0.0004

CAPITULO 12

DATOS DEL MARCO DE LA ENCUESTA DE POBLACIÓN ACTIVA

12.1. APLICACIÓN DEL MUESTREO SISTEMÁTICO CON INTERVALO VARIABLE A LOS DATOS DEL MARCO DE LA ENCUESTA DE POBLACIÓN ACTIVA[1]

En este estudio vamos a utilizar como marco para aplicar el programa POSDEM a la encuesta de población activa un listado de datos que se ha obtenido del cuaderno de tabulación de la sección 1-1 de San Sebastián. Se trata de una sección con cuatrocientas viviendas que ha sido seleccionada como parte de la muestra de secciones de la encuesta. En este capítulo no vamos a hacer referencia al hecho de tratarse de un muestreo en dos etapas. Esto es, esta sección es a su vez parte de una muestra de unidades de primera etapa, y que al realizar un nuevo muestreo dentro de la

[1] Este capítulo se ha mantenido con referencia a la versión DOS de POSDEM. El motivo ha sido la coherencia con los capítulos posteriores. No obstante un análisis de los datos con las últimas posibilidades disponibles en la versión Windows del programa POSDEM se incluyen como tablas de resultados para el País Vasco utilizando el marco del censo población de 1991.

sección estaríamos , en realidad, utilizando un diseño bietápico. Esta posibilidad se estudiará en el capítulo correspondiente.

Vamos a considerar , con carácter general que se trata de 400 viviendas que forman un estrato homogéneo próximos geográficamente. El fichero figura ordenado por número de miembros del hogar, y el número de identificación es un número correlativo. A efectos de seguir la explicación se reproduce aquí un listado con los hogares 72 a 82 del fichero mencionado. Estos datos se han incorporado a un fichero que se ha denominado EPA400.ASC.

EPA400.ASC

(1)	(2)
72	1
73	1
74	1
75	1
76	1
77	1
78	1
79	2
80	2
81	2
82	2

(1) Número correlativo de identificación de cada vivienda, ordenadas por número de personas que viven en ella.

(2) Valor de la variable para cada vivienda. Número de personas en cada vivienda.

Se puede observar que las viviendas están ordenadas por número de miembros del hogar. En los primeros 78 registros están aquellos hogares con un sólo miembro, después los hogares de dos miembros, etcétera ... Para hacer más rápido algunos cálculos en la

aplicación se ha incorporado un fichero que mantiene parecidas proporciones de composición de los hogares en cuanto a número de miembros pero que tiene solamente 100 registros. No obstante para el ejemplo que estamos siguiendo se va a trabajar con el fichero EPA400.ASC

Los pasos a seguir son los siguientes: desde el menú inicial de la aplicación se pulsa F10, de esta forma se puede definir el tamaño de población (400), el tamaño de muestra (20). Una vez fijados los valores de estos parámetros, se debe seleccionar uno de los ocho métodos propuestos. En este caso se opta por un esquema de selección sistemática estratificada con intervalo variable. El programa obtiene las 20 muestras posibles y realiza los cálculos necesarios para cada una de ellas. Hay que resaltar que si no se trata de un procedimiento sistemático las reiteraciones no estarían sujetas a ninguna condición, se pueden aumentar o disminuir. Una vez el programa concluye esta fase podemos comprobar cuales han sido los elementos seleccionados para las diferentes muestras.

```
        MUESTREO  SIST.ESTRATIF.SALT VARIA
======== MUESTRAS OBTENIDAS ==========================================
========( Matriz abreviada, Orden 1..L,1..M )=========================
=======================================================================
Nota: En filas se representa cada muestra obtenida, en columnas los valores
muestrales. Ejemplo el valor MUESTRA (4),U(2) representa a la unidad 2 de
la muestra 4

                U( 1)   U( 2)   U( 3)   U( 4)   U( 5)   U( 6)   U( 7)
 MUESTRA( 1)        1      22      43      64      85     106     127
 MUESTRA( 2)        2      23      44      65      86     107     128
 MUESTRA( 3)        3      24      45      66      87     108     129
 MUESTRA( 4)        4      25      46      67      88     109     130
 MUESTRA( 5)        5      26      47      68      89     110     131
 MUESTRA( 6)        6      27      48      69      90     111     132
 MUESTRA( 7)        7      28      49      70      91     112     133
 MUESTRA( 8)        8      29      50      71      92     113     134
 MUESTRA( 9)        9      30      51      72      93     114     135
 MUESTR( 10)       10      31      52      73      94     115     136
No se imprimen más de siete valores muestrales y diez muestras

 Muestras obtenidas =  28  Tamaño de muestra =  28  Tamaño de poblacion=  400
SELEC ... 1.- Menú  2.- Grafico ()  3.- Grafico =  4.- Listados  5.- Termina ( )
```

En este momento se puede definir cual es el fichero con los datos. Para esto se pulsa F4 y se escribe el nombre del fichero: EPA400.ASC . El programa tarda un poco en almacenar los datos puesto que tiene que casar la identificación de cada valor muestral con su correspondiente valor poblacional. Inmediatamente obtiene los datos para las especificaciones dadas. Se reproducen las pantallas correspondientes a los listados.

```
        MUESTREO  SIST.ESTRATIF.SALT VARIA
======== MUESTRAS OBTENIDAS ==========================================
========( Matriz abreviada, Orden 1..L,1..M )=========================
=====================================================================
Nota: En filas se representa cada muestra obtenida, en columnas los valores
muestrales. Ejemplo el valor MUESTRA (4),U(2) representa a la unidad 2 de
la muestra 4

                U( 1)   U( 2)   U( 3)   U( 4)   U( 5)   U( 6)   U( 7)
 MUESTRA( 1)      1       1       1       1       2       2       2
 MUESTRA( 2)      1       1       1       1       2       2       2
 MUESTRA( 3)      1       1       1       1       2       2       2
 MUESTRA( 4)      1       1       1       1       2       2       2
 MUESTRA( 5)      1       1       1       1       2       2       2
 MUESTRA( 6)      1       1       1       1       2       2       2
 MUESTRA( 7)      1       1       1       1       2       2       2
 MUESTRA( 8)      1       1       1       1       2       2       2
 MUESTRA( 9)      1       1       1       1       2       2       2
 MUESTR( 10)      1       1       1       1       2       2       2
No se imprimen más de siete valores muestrales y diez muestras

 Muestras obtenidas =  28  Tamaño de muestra =  28  Tamaño de poblacion=  400
SELEC ... 1.- Menú  2.- Grafico ()  3.- Grafico =  4.- Listados  5.- Termina ( )
```

Se puede comprobar la equivalencia entre las unidades que fueron seleccionadas con el procedimiento sistemático estratificado con intervalo variable y las que se obtienen al cusar estos datos con el fichero EPA400.ASC Así hasta la unidad 78 asigna el valor 1 que representa el número de miembros del hogar respectivo, y después sigue asignando a cada identificación el correspondiente valor de la variable para el fichero de datos con el que se está trabajando.

La opción (4) del menú de barra inferior nos conduce a un menú de listados, seleccionando la tercera opción, se obtiene un listado de esperanzas y varianzas que permite conocer que, para las especificaciones dadas, el error de muestreo en porcentaje es del 1.6%, y los límites de confianza para el estimador de la media son

3.03 y 3.23. Esto quiere decir que si obtenemos una muestra concreta que proporcione el valor 3.13 como número medio de personas por hogar, con ese error de muestreo podemos esperar que de cada 100 muestras que se obtienen en las mismas condiciones en 95 de ellas el valor estimado de la media estará comprendido entre 3.03 y 3.23. Si transcurrido un plazo de tiempo se vuelve a realizar la encuesta y, con un error similar, se obtiene un valor de 2.92 que presenta unos límites de confianza de (2.82 y 3.02), podemos asegurar que la variación es significativa y no se debe a una variación consecuencia de la variabilidad introducida por el proceso de muestreo.

```
==========================================================================
          MUESTREO  SIST.ESTRATIF.SALT VARIA
==========ESPERANZAS Y VARIANZAS PARA TODAS LAS MUESTRAS====================
==========================================================================
Nota: En filas se representa, la media,varianza, ... calculadas para
todas las muestras obtenidas. En columnas figura el estimador al que se
refieren los cálculos. Ejemplo, el valor media,varmed representa la media de
todas las varianzas del estimador media,calculada en base a todas las muestras
Atención: Estimadores sesgados de S y CVS, corregidos.
              MEDIA      VARMED      DESMED      LIMED       LSMED      CVSMED(%)
MEDIA        3.1325      0.0528      0.2299      2.6727      3.5923       7.3385
VARIANZA     0.0026      0.0009      0.0032      0.0002      0.0227       2.9687
DESVIACION   0.0507      0.0292      0.0568      0.0985      0.1588       1.7238
CO.VAR.(%)   1.6188     55.2622     25.4985      3.3689      4.2142      24.2861
LIM.SU.      3.2339      0.1112      0.3363      2.0688      3.0796      10.5486
LIM.IF.      3.0311     -0.0056      0.1092      2.5859      3.2764       3.6486
==========================================================================
==================== VALORES DE LA POBLACION ==========================
==========================================================================
              MEDIA      VARMED      DESMED      LIMED       LSMED      CVSMED(%)
VALORES      3.1325      0.1048      0.3225      2.4875      3.7775      10.2961
--------------------------------------------------------------------------
SELEC ... 1.- Menú  2.- Grafico <>  3.- Grafico =  4.- Listados  5.- Termina ( )
```

En este listado se puede disponer de las estimaciones de la media y de la varianza. Así, la esperanza de todas las estimaciones de las medias da el valor 3,13. La media de todas las varianzas es 0,0528. La varianza de las medias , el error de muestreo calculado en base a las reiteraciones, o error de muestreo que podemos denominar real para este experimento, es 0,0025.

En el próximo listado se pueden comprobar los resultados obtenidos para cada muestra al calcular la media de miembros por hogar y su error de muestreo en el caso de que la selección de la muestra se realice con la técnica de muestreo sistemático estratificado con intervalo variable, y que las estimaciones del error de muestreo se hayan corregido con la información proporcionada por las reiteraciones, para lo cual es necesario pulsar las teclas abreviadas Ctrl+T. Esto se puede comprobar, de una manera parcial, al analizar las estimaciones de la media del listado de cálculos para las diez primeras muestras.

MUESTREO SIST.ESTRATIF.SALT VARIA

======== Estimadores para cada muestra ========

Nota: En filas se representa cada muestra obtenida, en columnas los valores estimados. Ejemplo el valor MUESTRA(4),VARMED representa la varianza del estimador media calculado para la muestra 4

	MEDIA	VARMED	DESMED	LIMED	LSMED	CVSMED(%)
MUESTRA(1)	3.2500	0.0031	0.2883	2.6734	3.8266	8.8712
MUESTRA(2)	3.1500	0.0356	0.1887	2.7725	3.5275	5.9919
MUESTRA(3)	3.1000	0.0475	0.2179	2.6641	3.5359	7.0305
MUESTRA(4)	3.0500	0.0356	0.1887	2.6725	3.4275	6.1884
MUESTRA(5)	3.0500	0.0356	0.1887	2.6725	3.4275	6.1884
MUESTRA(6)	3.0500	0.0356	0.1887	2.6725	3.4275	6.1884
MUESTRA(7)	3.1000	0.0237	0.1541	2.7918	3.4082	4.9713
MUESTRA(8)	3.1000	0.0237	0.1541	2.7918	3.4082	4.9713
MUESTRA(9)	3.1500	0.0356	0.1887	2.7725	3.5275	5.9919
MUESTR(10)	3.1500	0.0356	0.1887	2.7725	3.5275	5.9919

No se imprimen más de diez muestras

SELEC ... 1.- Menú 2.- Grafico <> 3.- Grafico = 4.- Listados 5.- TerminaB(_)

12.2. COMPARACIÓN CON EL MUESTREO ALEATORIO SIMPLE SIN REPOSICIÓN

Con la opción de F6 se obtiene una pantalla con los resultados que se hubiesen podido obtener para los mismos datos pero aplicando muestreo aleatorio simple sin reposición. La pantalla muestra los valores calculados para la población de media y desviación típica, los valores de tamaño población y de muestra, y el error calculado con reiteraciones para el método sistemático estratificado con intervalo variable. También se puede estudiar el tamaño de muestra necesario en el muestreo sin reposición para distintos errores de muestreo en forma de coeficiente de variación porcentual.

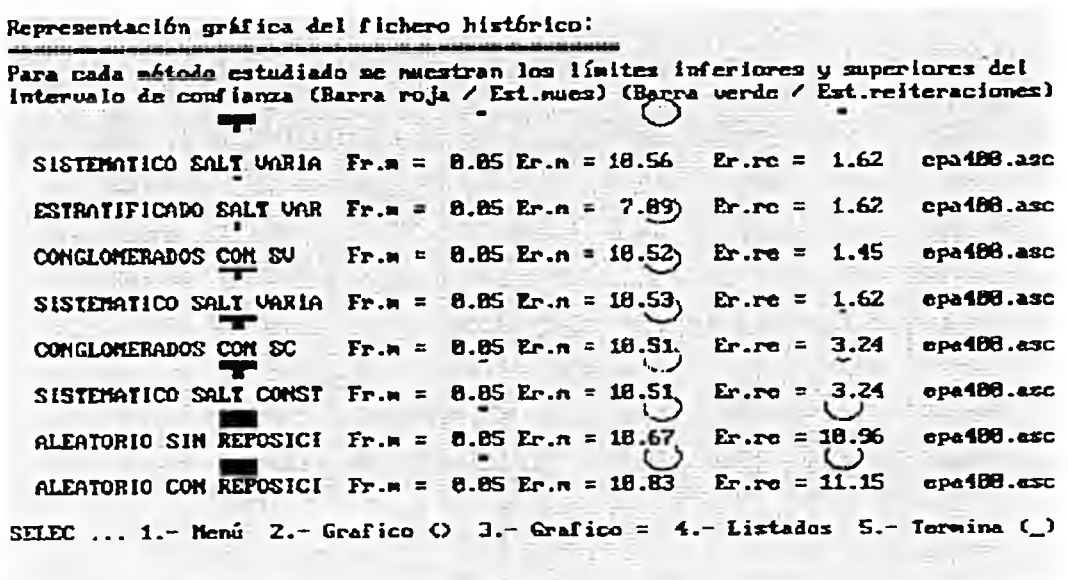
```
Modelo para comparar el tipo de muestreo  SIST.ESTRATIF.SALT VARIA
con el muestreo sin reposición .
═══════════════════════════════════════════════════════════════════
Media poblacional =  3.1325  Desviación típica =  1.479846
Tamaño de poblacion =  488  Tamaño de muestra =  20
Fracción de muestreo =  5 %  Error de muestreo 20% ( Real:
basado en las reiteraciones )=  1.617969 %
═══════════════════════════════════════════════════════════════════
El tamaño necesario en el muestreo sin reposición hubiese sido para
distintos coeficientes de variación ...
      El tamaño de muestra para media <- 20 %  =  6
      El tamaño de muestra para media <- 15 %  =  10
      El tamaño de muestra para media <- 10 %  =  21
      El tamaño de muestra para media <- 7 %  =  41
      El tamaño de muestra para media <- 5 %  =  73
      El tamaño de muestra para media <- 3 %  =  153
      El tamaño de muestra para media <- 2 %  =  233
      El tamaño de muestra para media <- 1 %  =  339
═══════════════════════════════════════════════════════════════════
Resultados para un error del  1.617969  el tamaño de muestra es  =  272
La ganancia en muestra (nsr-nn)/nsr es =  92.64706 %
SELEC ... 1.- Menú  2.- Grafico ()  3.- Grafico =  4.- Listados  5.- Termina (_)
```

Aquí se puede ver cómo para un tamaño de muestra de 21 viviendas, que es el que se ha determinado para este problema, obtendríamos un error de muestreo del 10 %. Y cómo si se fija el objetivo, independientemente del coste, de disminuir el error de muestreo hasta obtener un coeficiente de variación del 2% necesitaríamos un tamaño de muestra de 233 viviendas. Se vuelve a recordar que en el epígrafe anterior hemos comprobado que utilizando el método sistemático estratificado con intervalo variable se obtenía un error del 1.6 % para una muestra de 20 viviendas.

12.3. COMPARACIÓN CON OTROS MÉTODOS

Al efecto de comparar el método de muestreo sistemático estratificado con intervalo variable con otros métodos es posible guardar los principales resultados en un fichero que se denomina histórico. Para esto se pulsa la tecla de función F7, y después F8 si el fichero tiene información que se desea borrar, se puede realizar en este momento y volver a pulsar F7. Con F9 se visualiza este fichero con una representación gráfica.

Para obtener los datos con otro procedimiento se deben volver a ejecutar las secuencias F10, menú, F4, F7, y F9. Se proporciona una pantalla con los resultados para distintos métodos:

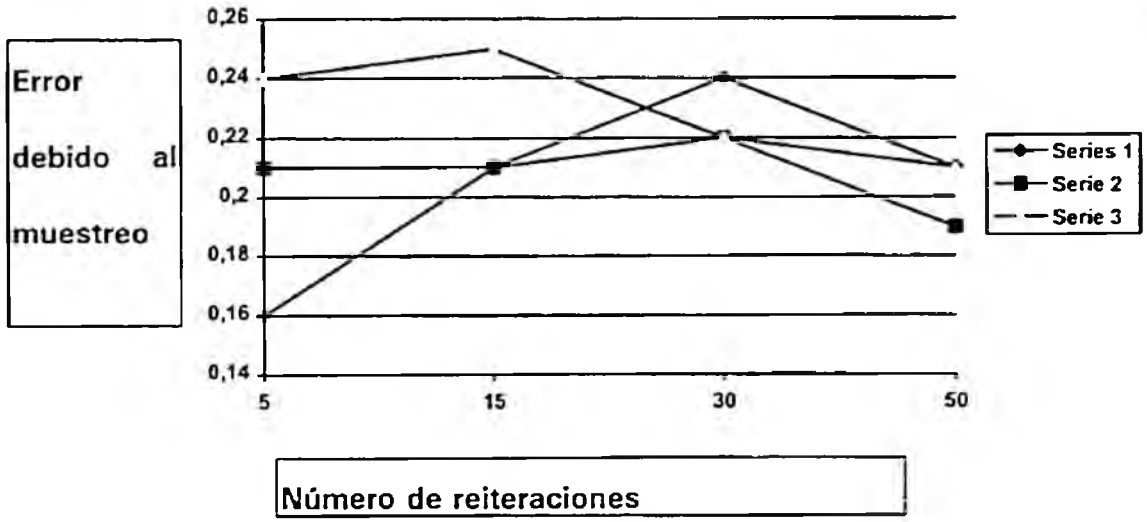


De este experimento se puede establecer que entre los métodos analizados el que presenta un óptimo en el proceso de selección es el muestreo sistemático estratificado con intervalo variable que permite reducir el error de muestreo basado en reiteraciones (Er.re.) del 11,15 al 1,62 y mejora también la estimación del error basada en valores muestrales (Er.m) que pasa de un 10,83 a un 7,09.

12.4. CONVERGENCIA DEL ESTIMADOR DE LA VARIANZA EL MUESTREO ALEATORIO SIMPLE SIN REPOSICIÓN

En el siguiente gráfico podemos comprobar cómo para el muestreo sin reposición a partir de 30 o 40 reiteraciones las estimaciones del error convergen hacia su verdadero valor. Las desviaciones que se aprecian no son significativas respecto del valor que se obtendría al calcular todas las muestras posibles. Por otra parte, esta precisión en cuanto al método sólo es necesaria para los métodos con o sin reposición, puesto que para los demás métodos que utilizan selección sistemática, sí se calcula todo el espacio muestral.

Gráfico 13.- Estimaciones del error de muestreo y número de reiteraciones



En este gráfico, con los datos de Epa100.asc, se puede observar la convergencia del estimador de la varianza al aumentar el número de reiteraciones para el muestreo sin reposición.

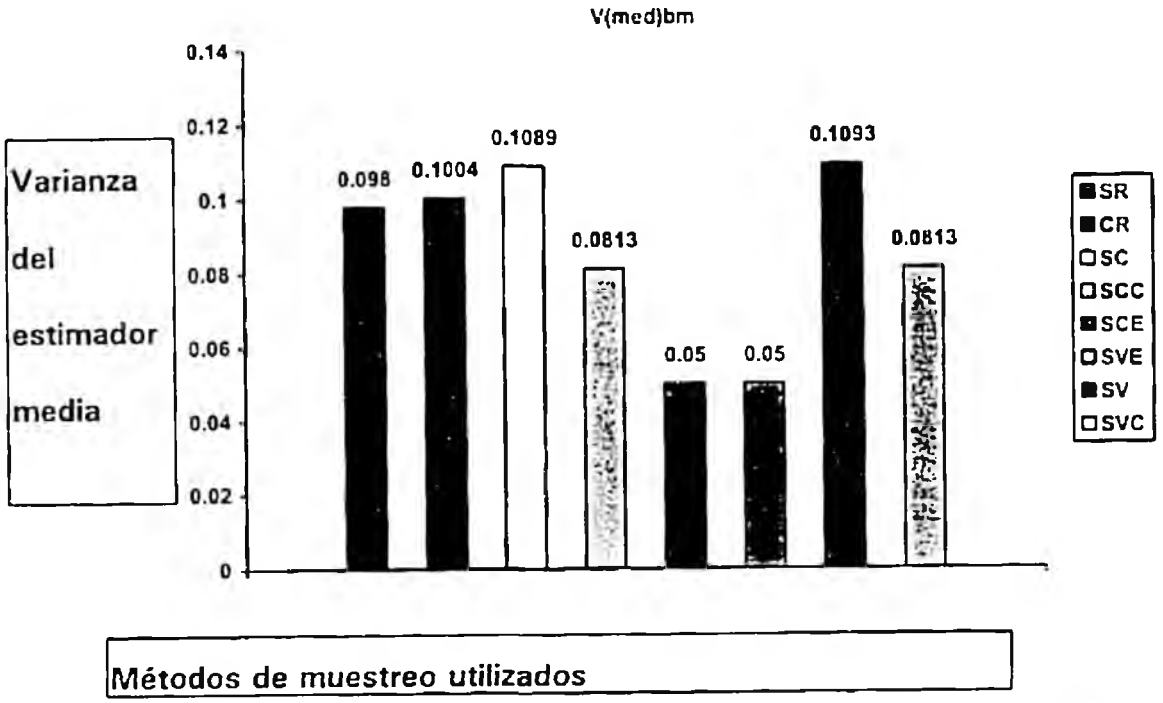
El verdadero valor es 0,21; este resultado se puede obtener ejecutando la aplicación para el fichero EPA100.ASC. Aquí se puede observar como para 4 reiteraciones este valor oscila entre 0,16 y 0,24 y como a medida que se aumenta el número de reiteraciones y se

repite el experimento los valores de la varianza se aproximan cada vez más a su verdadero valor.

12.5. UN MODELO DE COMPARACIÓN FACTORIAL

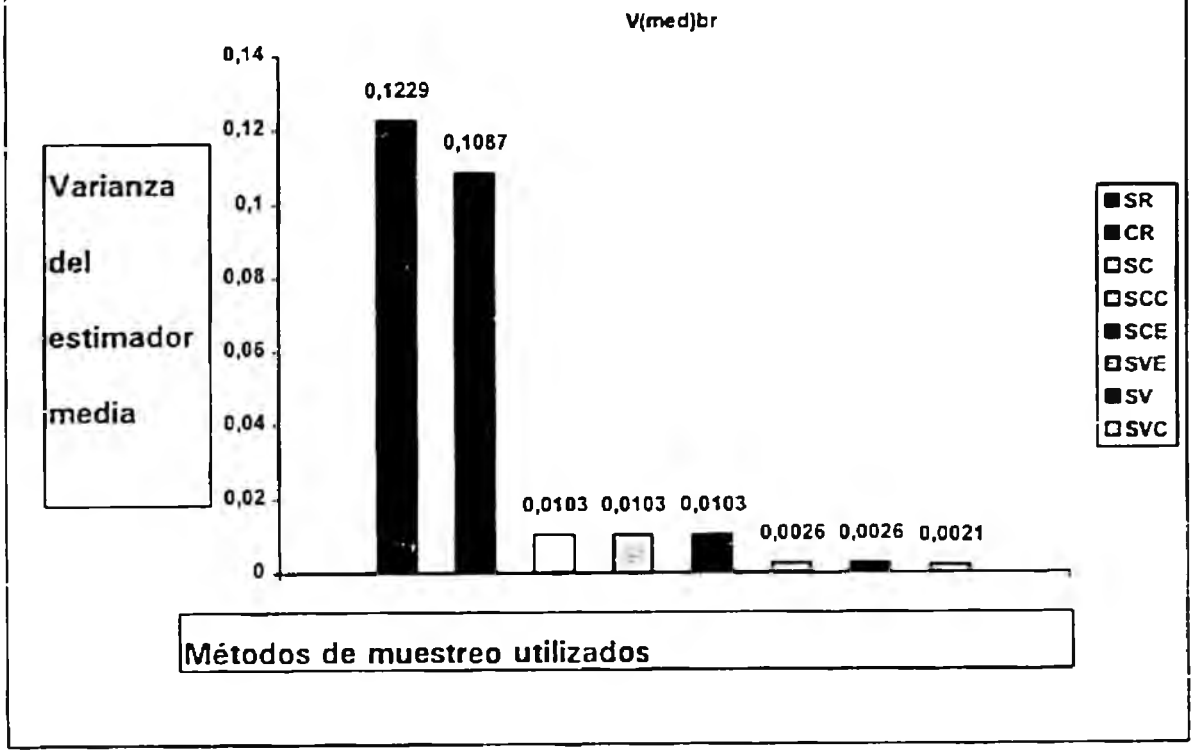
Continuando con el ejemplo de la sección de población activa a la que hemos aplicado los 8 métodos que permite la aplicación para probabilidades iguales, podemos obtener un gráfico de los errores de muestreo: varianza de la media basada en los valores muestrales, Varianza de la media basada en reiteraciones y Varianza de la varianza basada en las reiteraciones.

Gráfico 14.- Varianza estimada con datos muestrales.[2]



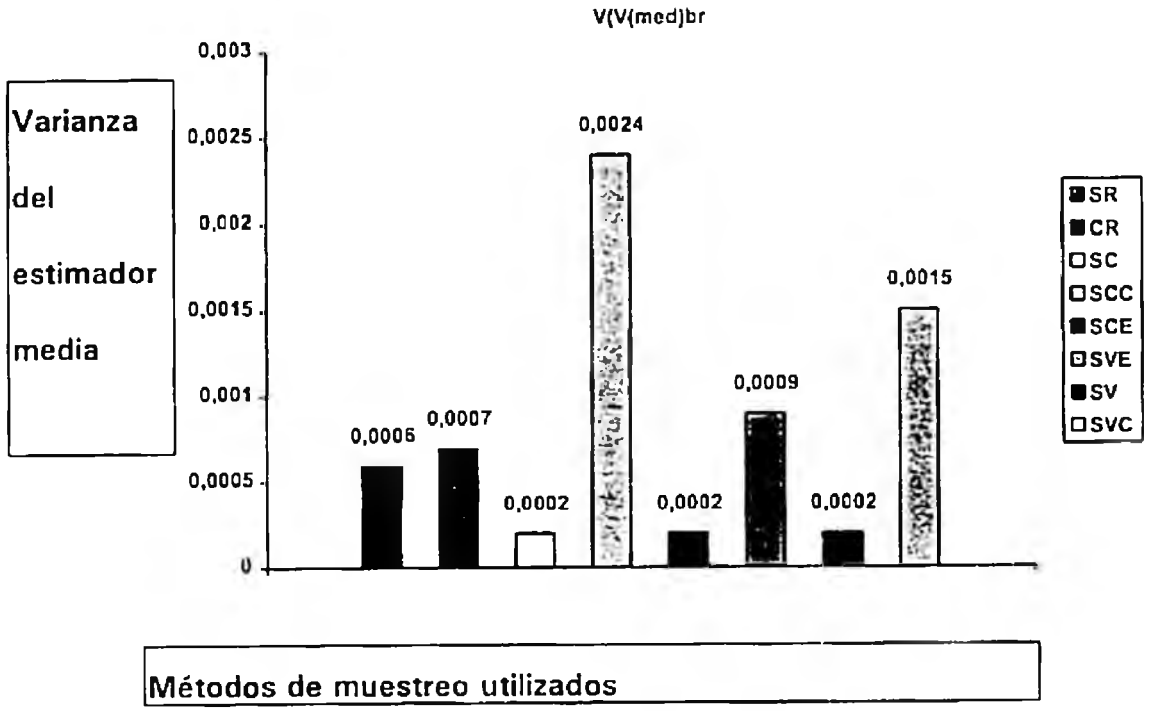
[2] SR Muestreo sin reposición.
CR Muestreo con reposición.
SC: Sistemático con intervalo constante.
SCC: Sistemático de conglomerados con intervalo constante.
SCE: Sistemático con intervalo constante y estimación del error con la técnica del lazo.
SVE: Sistemático con intervalo variable y estimación del error con la técnica del lazo.
SV: Sistemático con intervalo variable.
SVC: Sistemático con intervalo variable y estimación del error con la técnica del lazo.

Gráfico 15.- Varianza estimada con reiteraciones.[3]



[3] SR Muestreo sin reposición.
CR Muestreo con reposición.
SC: Sistemático con intervalo constante.
SCC: Sistemático de conglomerados con intervalo constante.
SCE: Sistemático con intervalo constante y estimación del error con la técnica del lazo.
SVE: Sistemático con intervalo variable y estimación del error con la técnica del lazo.
SV: Sistemático con intervalo variable.
SVC: Sistemático con intervalo variable y estimación del error con la técnica del lazo.

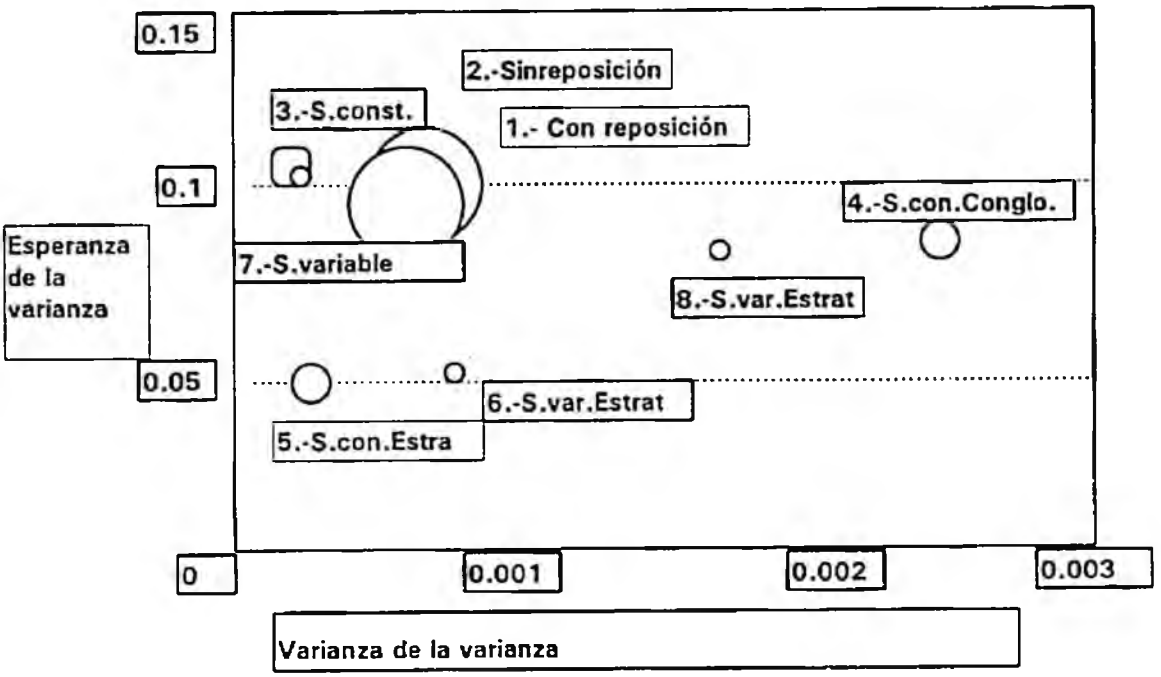
Gráfico 16.- Varianza del estimador de la varianza basado en reiteraciones.[4]



[4] SR Muestreo sin reposición.
CR Muestreo con reposición.
SC: Sistemático con intervalo constante.
SCC: Sistemático de conglomerados con intervalo constante.
SCE: Sistemático con intervalo constante y estimación del error con la técnica del lazo.
SVE: Sistemático con intervalo variable y estimación del error con la técnica del lazo.
SV: Sistemático con intervalo variable.
SVC: Sistemático con intervalo variable y estimación del error con la técnica del lazo.

Si numeramos los métodos de mayor a menor error de muestreo calculado en base a reiteraciones. El método que presenta un menor error de muestreo, basado en reiteraciones, es el muestreo de conglomerado con intervalo variable (0,0021); los métodos que presentan un menor valor esperado del error, esto es la esperanza de la varianza del estimador media obtenida en base a la información de cada muestra, son el muestreo estratificado con intervalo variable o con intervalo constante (0,05); y, por último, que los métodos que presenta una menor varianza de la varianza son los sistemáticos con intervalo variable o constante y el constante con estratificación (0,0002). Para poder observar mejor estos resultados se ha incluido un gráfico que recoge esta información.

Gráfico 17.- Comparación de varianzas para los datos analizados del marco de la Encuesta de Población Activa.[5]



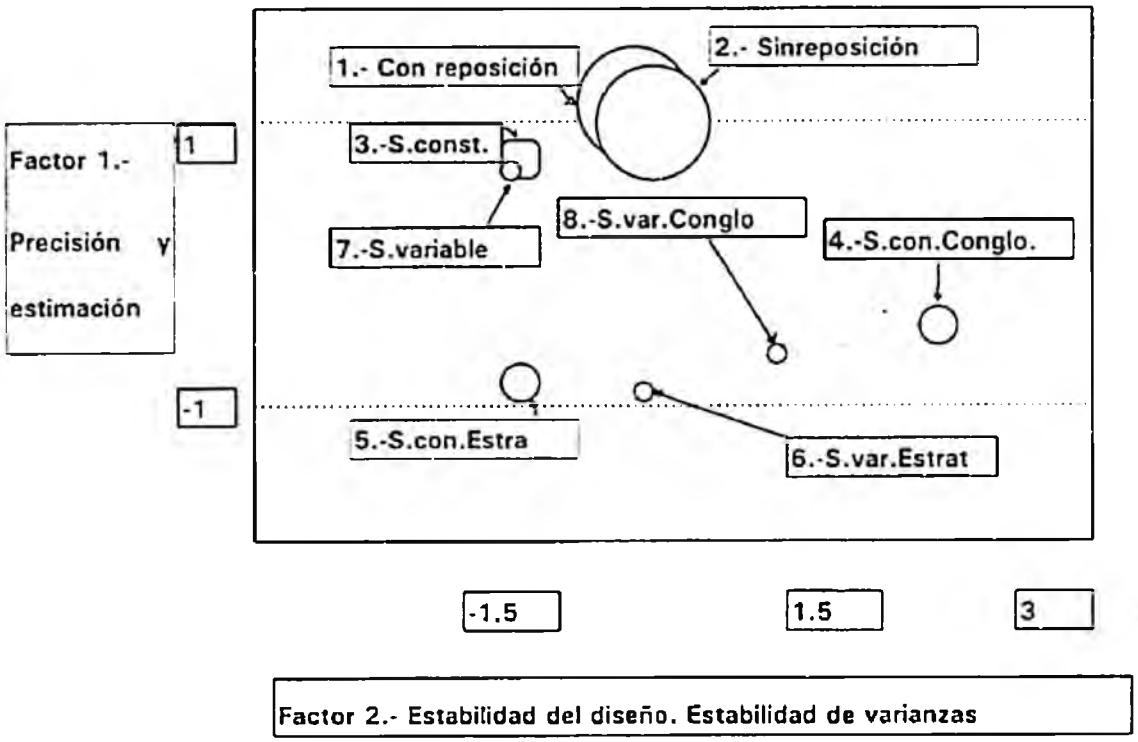
[5] La burbuja es proporcionale al error basado en reiteraciones.

Si el criterio para seleccionar un método u otro es que el error de muestreo real sea el más pequeño posible, el método elegido será el 8; ahora bien, si el criterio es minimizar cualquiera de las otras dos variables, los métodos elegidos serán otros. Para poder reducir la información suministrada por las tres variables consideradas es posible realizar un análisis factorial, que permita tener una idea espacial en un eje de dos dimensiones. Los resultados del análisis se proporcionan al final del capítulo. Aquí se va a reproducir un gráfico con los dos factores obtenidos para los ocho métodos.

Tabla 28.- Modelo factorial: precisión, estimación y estabilidad del diseño.

D_R Metodo	Vmed	Vmedr	Vvr	Fact1	Fact2
1.-CR	.09	.1229	.0006	1.40	-.094
2.-SR	.10	.1087	.0007	1.30	.001
3.-SC	.10	.0103	.0002	.21	-.893
4.-SCC	.08	.0103	.0024	-.17	1.987
5.-SCE	.05	.0103	.0002	-1.23	-.887
6.-SV	.05	.0026	.0009	-1.24	.009
7.-SV	.10	.0026	.0002	.12	-.912
8.-SVC	.08	.0021	.0015	-.39	.790

Gráfico 18.- Modelo factorial: precisión, estimación y estabilidad del diseño.



En este gráfico podemos comprobar como se sitúan los diferentes métodos en relación a los dos ejes factoriales. Del estudio de estos factores se puede deducir que el primero, el situado en el eje de ordenadas, recoge la información referida al error de muestreo, tanto si es el obtenido por la información de las muestras o directamente por las reiteraciones; el segundo factor, en el eje de abcisas recoge fundamentalmente el efecto de los métodos sobre la variación de la varianza, recoge por tanto el efecto de estabilidad del diseño. Así cuanto más arriba, en el eje de ordenadas , esté situado el método, será menos preciso; y cuanto más a la derecha, en el eje de abcisas, el método será menos estable. Por tanto, serán más interesantes los métodos que aparezcan en el cuadrante inferior izquierdo, que denotarán precisión del estimador y estabilidad de su varianza. En este caso, se puede elegir entre los métodos estratificado con intervalo variable o con intervalo constante, que son los mejor situados. Para poder tener una idea más completa se ha incluido en el gráfico la información relativa al error real de muestreo de forma que las burbujas sean más o menos grandes según que el error sea más o menos grande. De esta forma podemos apreciar que el método de intervalo variable si bien pierde estabilidad tiene un error

de muestreo significativamente inferior al método de intervalo constante. Una solución que mejoraría mucho el procedimiento de intervalo variable sería el disponer de un método de estimación del error, distinto al de unir dos unidades por estrato, de forma que el método fuese más estable respecto de su error.

Resultados del análisis de factores.

Initial Factor Method: Principal Components
Prior Communality Estimates: ONE
Eigenvalues of the Correlation Matrix: Total = 3 Average = 1

	1	2	3
Eigenvalue	1.479891	0.885197	0.634911
Difference	0.594694	0.250286	
Proportion	0.4933	0.2951	0.2116
Cumulative	0.4933	0.7884	1.0000

2 factors will be retained by the NFACTOR criterion.
Factor Pattern

	FACTOR1	FACTOR2
VMED	0.78236	0.23488
VMEDR	0.75996	0.35348
VVR	-0.53875	0.83969

Variance explained by each factor

FACTOR1	FACTOR2
1.479891	0.885197

Final Communality Estimates: Total = 2.365089

VMED	VMEDR	VVR
0.667262	0.702486	0.995340

Rotation Method: Equamax
Orthogonal Transformation Matrix

	1	2
1	0.89194	-0.45216
2	0.45216	0.89194

Rotated Factor Pattern

	FACTOR1	FACTOR2
VMED	0.80402	-0.14426
VMEDR	0.83767	-0.02835
VVR	-0.10086	0.99256

Variance explained by each factor

FACTOR1	FACTOR2
1.358305	1.006783

Final Communality Estimates: Total = 2.365089

VMED	VMEDR	VVR
0.667262	0.702486	0.995340

Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with each Factor

FACTOR1	FACTOR2
1.000000	1.000000

Standardized Scoring Coefficients

	FACTOR1	FACTOR2
VMED	0.59151	-0.00238
VMEDR	0.63859	0.12397
VVR	0.10421	1.01069

Initial Factor Method: Principal Components
Prior Communality Estimates: ONE
Eigenvalues of the Correlation Matrix: Total = 3 Average = 1

	1	2	3
Eigenvalue	1.479891	0.885197	0.634911
Difference	0.594694	0.250286	
Proportion	0.4933	0.2951	0.2116
Cumulative	0.4933	0.7884	1.0000

2 factors will be retained by the NFACTOR criterion.

Factor Pattern

	FACTOR1	FACTOR2
VMED	0.78236	0.23488
VMEDR	0.75996	0.35348
VVR	-0.53875	0.83969

Variance explained by each factor

FACTOR1	FACTOR2
1.479891	0.885197

Final Communality Estimates: Total = 2.365089

VMED	VMEDR	VVR
0.667262	0.702486	0.995340

Rotation Method: Equamax
Orthogonal Transformation Matrix

	1	2
1	0.89194	-0.45216
2	0.45216	0.89194

Rotated Factor Pattern

	FACTOR1	FACTOR2
VMED	0.80402	-0.14426
VMEDR	0.83767	-0.02835
VVR	-0.10086	0.99256

Variance explained by each factor

FACTOR1	FACTOR2
1.358305	1.006783

Final Communality Estimates: Total = 2.365089

VMED	VMEDR	VVR
0.667262	0.702486	0.995340

Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with each Factor

FACTOR1	FACTOR2
1.000000	1.000000

Standardized Scoring Coefficients

	FACTOR1	FACTOR2
VMED	0.59151	-0.00238
VMEDR	0.63859	0.12397
VVR	0.10421	1.01069

Tabla 29.- Resultados para el conjunto del País Vasco utilizando el marco del censo de 1991.

	Activos	Ocupados	Parados
Estimador del total de parados [EPA (II TR /91)]	862.300	698.900	163.400
Varianza del estimador [EPA (II TR /91)]	135,5E+7	129,8E+7	528,6E+6
Desviación del estimador [EPA (II TR /91)]	11.641	11.392	7.141
Coeficiente de variación % [EPA (II TR /91)]	1,35%	1,63%	4,37%
Total de parados según censo [CERCA]	868.954	700.042	168.912
Tamaños de población y muestra	N=1600	n=160	
Métodos de muestreo	*Desviaciones del estimador*		
Sin reposicion y probabilidades iguales	25.273	21.202	5.810
Estratificado con una unidad por estrato	6.898	7.506	3.657
Sistemático con intervalo constante	8.046	6.332	2.138
Sistemático equilibrado	8.067	7.597	2.853
Sistemático modificado	4.065	4.302	2.107
Sistemático con intervalo variable	7.567	6.867	3.077
Centrado con intervalo constante	6.905	3.870	2.020
Centrado con intervalo variable	6.680	5.295	2.120
Con reposición y probabilidades proporcionales	6.446	7.983	3.428
Sin reposición y probabilidades proporcionales	6.878	7.443	2.568
Con reposición parcial y probabilidades proporcionales	6.537	7.534	3.600
	Coeficiente de variación		
Sin reposicion y probabilidades iguales	2,91%	3,03%	3,44%
Estratificado con una unidad por estrato	0,79%	1,07%	2,17%
Sistemático con intervalo constante	0,93%	0,90%	1,27%
Sistemático equilibrado	0,93%	1,09%	1,69%
Sistemático modificado	0,47%	0,61%	1,25%
Sistemático con intervalo variable	0,87%	0,98%	1,82%
Centrado con intervalo constante	0,79%	0,55%	1,20%
Centrado con intervalo variable	0,77%	0,76%	1,26%
Con reposición y probabilidades proporcionales	0,74%	1,14%	2,03%
Sin reposición y probabilidades proporcionales	0,79%	1,06%	1,52%
Con reposición parcial y probabilidades proporcionales	0,75%	1,08%	2,13%

Tabla 30.- Resultados para la provincia de Vizcaya utilizando el marco del censo de 1991.

	Activos	Ocupados	Parados
Estimador del total de parados[EPA (II TR /91)]	473.300	371.900	101.400
Varianza del estimador [EPA (II TR /91)]	123,7E+7	110,0E+7	325,9E+6
Desviación del estimador [EPA (II TR /91)]	11.123	10.488	5.709
Coeficiente de variación % [EPA (II TR /91)]	2,35%	2,82%	5,63%
Total de parados según censo [CERCA]	467.951	370.355	97.596
Tamaños de población y muestra	N=900	n=75	
Métodos de muestreo	*Desviaciones del estimador*		
Sin reposicion y probabilidades iguales	19.608	16.189	4.795
Estratificado con una unidad por estrato	5.121	5.616	2.971
Sistemático con intervalo constante	7.161	7.962	2.514
Sistemático equilibrado	4.304	5.088	2.233
Sistemático modificado	4.642	5.762	3.320
Sistemático con intervalo variable	3.170	3.114	2.432
Centrado con intervalo constante	7.091	3.290	3.801
Centrado con intervalo variable	1.405	766	639
Con reposición y probabilidades proporcionales	5.494	6.428	2.731
Sin reposición y probabilidades proporcionales	5.391	5.431	2.665
Con reposición parcial y probabilidades proporcionales	4.983	5.344	2.926
	Coeficiente de variación		
Sin reposicion y probabilidades iguales	4,19%	4,37%	4,91%
Estratificado con una unidad por estrato	1,09%	1,52%	3,04%
Sistemático con intervalo constante	1,53%	2,15%	2,58%
Sistemático equilibrado	0,92%	1,37%	2,29%
Sistemático modificado	0,99%	1,56%	3,40%
Sistemático con intervalo variable	0,68%	0,84%	2,49%
Centrado con intervalo constante	1,52%	0,89%	3,89%
Centrado con intervalo variable	0,30%	0,21%	0,65%
Con reposición y probabilidades proporcionales	1,17%	1,74%	2,80%
Sin reposición y probabilidades proporcionales	1,15%	1,47%	2,73%
Con reposición parcial y probabilidades proporcionales	1,06%	1,44%	3,00%

Tabla 31.- Resultados para la provincia de Guipúzcoa utilizando el marco del censo de 1991.

	Activos	Ocupados	Parados
Estimador del total de parados [EPA (II TR /91)]	278.800	233.600	45.200
Varianza del estimador [EPA (II TR /91)]	411,2E+6	368,9E+6	844,7E+5
Desviación del estimador [EPA (II TR /91)]	6.412	6.074	2.906
Coeficiente de variación % [EPA (II TR /91)]	2,30%	2,60%	6,43%
Total de parados según censo [CERCA]	283.988	232.392	51.596
Tamaños de población y muestra	N=480	n=60	
Métodos de muestreo	*Desviaciones del estimador*		
Sin reposicion y probabilidades iguales	12.523	10.525	2.795
Estratificado con una unidad por estrato	3.290	3.407	1.825
Sistemático con intervalo constante	4.115	4.611	1.207
Sistemático equilibrado	3.458	3.134	2.048
Sistemático modificado	3.738	3.616	1.447
Sistemático con intervalo variable	2.945	2.478	1.999
Centrado con intervalo constante	2.688	1.228	1.460
Centrado con intervalo variable	520	572	1.092
Con reposición y probabilidades proporcionales	3.187	2.764	1.932
Sin reposición y probabilidades proporcionales	3.490	3.256	1.735
Con reposición parcial y probabilidades proporcionales	3.095	2.996	2.043
	Coeficiente de variación		
Sin reposicion y probabilidades iguales	4,41%	4,53%	5,42%
Estratificado con una unidad por estrato	1,16%	1,47%	3,54%
Sistemático con intervalo constante	1,45%	1,98%	2,34%
Sistemático equilibrado	1,22%	1,35%	3,97%
Sistemático modificado	1,32%	1,56%	2,80%
Sistemático con intervalo variable	1,04%	1,07%	3,87%
Centrado con intervalo constante	0,95%	0,53%	2,83%
Centrado con intervalo variable	0,18%	0,25%	2,12%
Con reposición y probabilidades proporcionales	1,12%	1,19%	3,75%
Sin reposición y probabilidades proporcionales	1,23%	1,40%	3,36%
Con reposición parcial y probabilidades proporcionales	1,09%	1,29%	3,96%

Tabla 32.- Resultados para la provincia de Alava utilizando el marco del censo de 1991.

	Activos	Ocupados	Parados
Estimador del total de parados [EPA (II TR /91)]	110.200	93.300	16.900
Varianza del estimador [EPA (II TR /91)]	128,3E+6	145,6E+6	405,1E+5
Desviación del estimador[EPA (II TR /91)]	3.582	3.816	2.013
Coeficiente de variación % [EPA (II TR /91)]	3,25%	4,09%	11,91%
Total de parados según censo [CERCA]	116.703	97.005	19.698
Tamaños de población y muestra	N=220	n=44	
Métodos de muestreo	*Desviaciones del estimador*		
Sin reposicion y probabilidades iguales	7.171	6.086	1.439
Estratificado con una unidad por estrato	2.136	2.222	837
Sistemático con intervalo constante	3.074	3.184	713
Sistemático equilibrado	2.171	1.986	550
Sistemático modificado	2.371	1.982	521
Sistemático con intervalo variable	1.670	1.858	593
Centrado con intervalo constante	3.376	3.599	842
Centrado con intervalo variable	1.148	1.751	731
Con reposición y probabilidades proporcionales	2.254	2.575	888
Sin reposición y probabilidades proporcionales	2.324	2.352	953
Con reposición parcial y probabilidades proporcionales	2.275	1.819	857
	Coeficiente de variación		
Sin reposicion y probabilidades iguales	6,14%	6,27%	7,31%
Estratificado con una unidad por estrato	1,83%	2,29%	4,25%
Sistemático con intervalo constante	2,63%	3,28%	3,62%
Sistemático equilibrado	1,86%	2,05%	2,79%
Sistemático modificado	2,03%	2,04%	2,64%
Sistemático con intervalo variable	1,43%	1,92%	3,01%
Centrado con intervalo constante	2,89%	3,71%	4,28%
Centrado con intervalo variable	0,98%	1,81%	3,71%
Con reposición y probabilidades proporcionales	1,93%	2,65%	4,51%
Sin reposición y probabilidades proporcionales	1,99%	2,42%	4,84%
Con reposición parcial y probabilidades proporcionales	1,95%	1,88%	4,35%

CAPÍTULO 13

DATOS DEL MARCO DE LA ENCUESTA INDUSTRIAL Y OTROS

13.1. APLICACIÓN DEL MUESTREO SISTEMÁTICO VARIABLE A LOS DATOS DEL MARCO DE LA ENCUESTA INDUSTRIAL[1]

El marco que se va a utilizar es un listado que contiene los datos relativos a 160 empresas de un determinado estrato , que, por ejemplo, pueden ser del sector de componentes electrónicos , con sede en una determinada zona y con un tamaño, medido en número de personas contratadas, que varía entre 20 y 170 trabajadores. El estrato se ha definido de manera que es lo suficientemente homogéneo en su interior como para que pueda llevarse a cabo un proceso de muestreo. En el diseño de las encuestas económicas existe la práctica de no utilizar muestreo para empresas con más de veinte trabajadores , no obstante en este caso se va a realizar el estudio para ese estrato. La práctica de disponer de datos para el

[1] En este capítulo se utiliza la versión DOS de POSDEM por coherencia con los capítulos anterior y posterior.

conjunto de la población ha llevado, en este estudio, a fijar un tamaño de muestra bastante elevado, una fracción de muestreo del 25%, precisamente para disponer de intervalos de confianza lo más ajustados posibles, que eviten , en lo posible , el inconveniente de sustituir datos de una población por los de una muestra. Por este mismo motivo se hace especial hincapié en un apartado posterior sobre la detección de unidades que deben formar parte de la muestra con probabilidad uno, esto es, en la detección de empresas que estarán en el estudio con certeza, empresas autorepresentadas.

Por ahora los datos del problema son: el método de selección sistemático con intervalo variable, la población es de 160 empresas, el tamaño de muestra es igual a 40 unidades, y el espacio muestral está formado por las 4 muestras posibles. Una vez introducidos los datos de identificación de cada empresa y el número de trabajadores que tiene cada empresa en un fichero que se ha denominado EIE160.ASC y, seguidos los pasos necesarios se llega a obtener un listado de esperanzas y varianzas como el mostrado a continuación.

MUESTREO SIST.ESTRATIF.SALT VARIA

ESPERANZAS Y VARIANZAS PARA TODAS LAS MUESTRAS

Nota: En filas se representa, la media,varianza, ... calculadas para todas las muestras obtenidas. En columnas figura el estimador al que se refieren los cálculos. Ejemplo, el valor media,varmed representa la media de todas las varianzas del estimador media,calculada en base a todas las muestras

Atención: Estimadores sesgados de S y CVS, corregidos.

	MEDIA	VARMED	DESMED	LIMED	LSMED	CVSMED(%)
MEDIA	54.6563	5.9884	2.4291	49.7981	59.5144	4.4443
VARIANZA	8.8381	5.3137	8.2111	1.1664	8.5825	8.7389
DESVIACION	8.1735	2.3851	8.4594	1.0886	8.7632	8.8549
CO.VAR.(%)	8.3175	39.8677	19.2619	2.1649	1.2843	19.5787
LIM.SU.	55.8833	18.5187	3.3841	52.8458	68.9532	6.8764
LIM.IF.	54.3892	1.2981	1.4663	47.7258	57.9882	2.6567

VALORES DE LA POBLACION

	MEDIA	VARMED	DESMED	LIMED	LSMED	CVSMED(%)
VALORES	54.6563	29.7232	5.4519	43.7524	65.5601	9.9749

SELEC ... 1.- Menú 2.- Grafico <> 3.- Grafico = 4.- Listados 5.- Termina (_)

De este listado se deduce que el error de muestreo, el coeficiente de variación del estimador en porcentaje y calculado en base a todo el espacio muestral, es 0.3171 %. Por lo que los límites inferior y superior de confianza al 95 % son respectivamente 54.3 y 55.0 .

Esto quiere decir que, si transcurrido un cierto plazo, se repite la investigación y se observa, con un error de muestreo similar, que el estimador de la media es 53.8, con unos límites entre 53.4 y 54.2, la caída apreciada desde el nivel 54.6 a 53.8, es estadísticamente significativa y no ha sido consecuencia de la

variabilidad introducida por el proceso de muestreo, por utilizar una muestra y no la población completa.

Un listado de los cálculos para cada muestra , una vez corregida la estimación del error con la información suministrada por las reiteraciones al pulsar la tecla abreviada ctrl+T, permite comprobar como los límites de confianza establecidos se mantienen para cada una de las cuatro posibilidades del espacio muestral.

```
        MUESTREO  SIST.ESTRATIF.SALT VARIA
======== Estimadores para cada muestra ==========================
==========================================================
==========================================================

Nota: En filas se representa cada muestra obtenida, en columnas los valores
estimados. Ejemplo el valor MUESTRA(4),VARMED representa la varianza del
estimador media calculado para la muestra 4

                MEDIA      VARMED      DESMED      LIMED      LSMED     CVSMED(%)
MUESTRA( 1)    54.8500      4.6687      2.1607    50.5285    59.1715        3.9393
MUESTRA( 2)    54.6000      5.9156      2.4322    49.7356    59.4644        4.4546
MUESTRA( 3)    54.4000      9.5986      3.0969    48.2863    60.5938        5.6928
MUESTRA( 4)    54.7750      3.4266      1.8511    51.0728    58.4772        3.3795

Muestras obtenidas =  4  Tamaño de muestra =  40  Tamaño de poblacion=  160


SELEC ... 1.- Menú  2.- Grafico <>  3.- Grafico =  4.- Listados  5.- Termina (_)
```

Esto es, sea cual sea, la muestra finalmente seleccionada los valores del estimador media variaran entre el 54.4 de la tercera muestra y 54.8 de la primera.

13.2. COMPARACIÓN CON EL MUESTREO ALEATORIO SIMPLE SIN REPOSICIÓN

Se puede realizar una comparación de estos resultados con los que se hubiesen conseguido en el caso de aplicar un método de selección de muestras aleatorias sin reposición, mediante la siguiente pantalla la cual se obtiene al pulsar la tecla de función F6.

```
Modelo para comparar el tipo de muestreo  SIST.ESTRATIF.SALT VARIA
con el muestreo sin reposición .
═══════════════════════════════════════════════════════════════════
Media poblacional =  54.65625  Desviación típica =  39.81585
Tamaño de poblacion =  168  Tamaño de muestra =  40
Fracción de muestreo =  25 %  Error de muestreo CV% ( Real:
basado en las reiteraciones )=  .3175162 %
═══════════════════════════════════════════════════════════════════
El tamaño necesario en el muestreo sin reposición hubiese sido para
distintos coeficientes de variación ...
        El tamaño de muestra para media ← 20  %  =  12
        El tamaño de muestra para media ← 15 %  =  21
        El tamaño de muestra para media ← 10 %  =  40
        El tamaño de muestra para media ← 7 %  =  65
        El tamaño de muestra para media ← 5 %  =  91
        El tamaño de muestra para media ← 3 %  =  126
        El tamaño de muestra para media ← 2 %  =  143
        El tamaño de muestra para media ← 1 %  =  155
═══════════════════════════════════════════════════════════════════
 Resultados para un error del  .3175162  el tamaño de muestra es  =  168
 La ganancia en muestra (nsr-nn)/nsr es =  75 %
SELEC ... 1.- Menú  2.- Grafico <>  3.- Grafico =  4.- Listados  5.- Termina ( )
```

En esta pantalla se puede comprobar que, con la misma población, una muestra de 40 empresas y un procedimiento de

selección aleatorio sin reposición , el error de muestreo, en términos de coeficiente de variación del estimador en porcentaje, toma un valor en torno al 10%. Se recuerda que con el procedimiento de selección sistemático estratificado con intervalo variable el error, en los mismos términos, no llega al 1%, pues, se sitúa en el 0.3%

Otra forma de analizar la misma información es señalando que, si se utiliza un muestreo sin reposición y se fija el error de muestreo en el 1%, la investigación requerirá un tamaño de muestra de 155 unidades que, al ser la población del estrato igual a 160, convierte en realidad la encuesta en un censo.

13.3. COMPARACIÓN CON OTROS MÉTODOS

Sin más que repetir los pasos definidos para cada método es posible llegar a la siguiente tabla donde se muestran los resultados para el error de muestreo según el método elegido.

```
Representación gráfica del fichero histórico:
**********************************************
Para cada método estudiado se muestran los límites inferiores y superiores del
intervalo de confianza (Barra roja / Est.mues) (Barra verde / Est.reiteraciones)

 CONGLOMERADOS SIST. SV  Pr.m =  0.25 Er.m =  6.90   Er.re =  0.70   eie160.asc
 CONGLOMERADOS SIST. SC  Pr.m =  0.25 Er.m =  6.54   Er.re =  2.11   eie160.asc
 SIST.ESTRATIF.SALT CON  Pr.m =  0.25 Er.m =  4.18   Er.re =  2.11   eie160.asc
 SISTEMATICO SALT VARIA  Pr.m =  0.25 Er.m = 10.10   Er.re =  0.32   eie160.asc
 SISTEMATICO SALT CONST  Pr.m =  0.25 Er.m = 10.10   Er.re =  2.11   eie160.asc
 ALEATORIO CON REPOSICI  Pr.m =  0.25 Er.m = 12.23   Er.re =  9.12   eie160.asc
 SIST.ESTRATIF.SALT VAR  Pr.m =  0.25 Er.m =  4.44   Er.re =  0.32   eie160.asc
SELEC ... 1.- Menú  2.- Grafico <>  3.- Grafico =  4.- Listados  5.- Termina (_)
```

Para estos datos el método que presenta menor error de muestreo real es el sistemático estratificado con intervalo variable. No obstante, para comprender mejor los resultados pueden extraerse de la aplicación los datos necesarios para realizar el análisis gráfico que se realiza a continuación . Con el único fin de utilizar un fichero más

reducido que disminuya el tiempo de espera, se ha incorporado a la aplicación un fichero similar al utilizado anteriormente pero con la particularidad de tener únicamente 40 empresas. Es a partir de estos datos de donde se han extraído las cifras que se manejan en el resto del capítulo.

Gráfico 19.- Comparación de varianzas para los datos del marco de la Encuesta Industrial. La burbuja es proporcionales al error basado en reiteraciones.

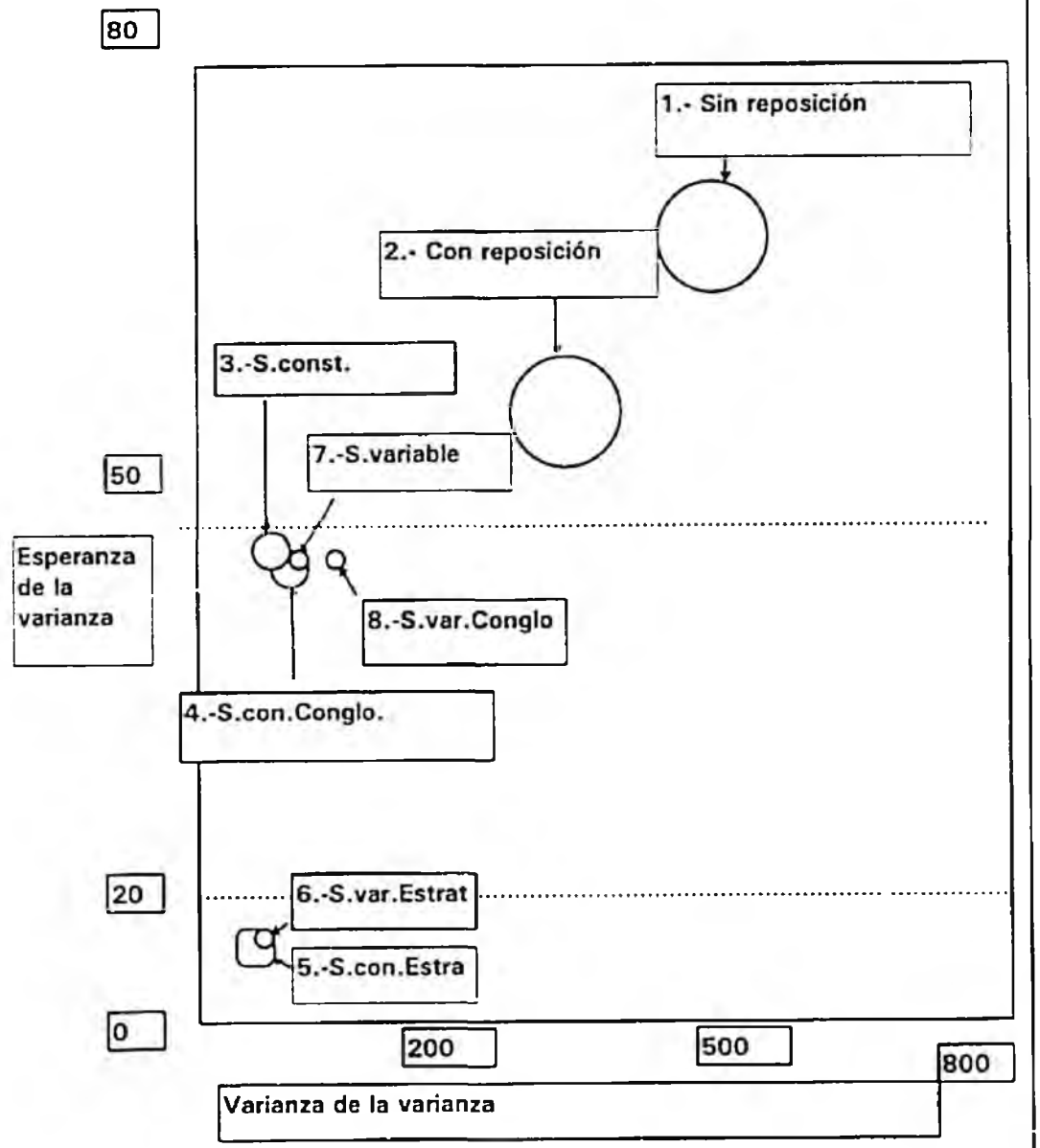


Tabla 33.- Varianzas y estabilidad para los datos del marco de la Encuesta Industrial

Método	Vmed	Vmedr	Vvr
1.-SR	51.48	48.87	503.52
2.-CR	40.51	31.92	282.09
3.-SC	42.82	7.56	7.04
4.-SCC	40.55	7.56	22.23
5.-SCE	8.48	7.56	1.35
6.-SVE	9.82	0.67	5.27
7.-SV	43.61	0.67	44.03
8.-SVC	43.19	0.71	104.43

En este experimento el análisis factorial no proporciona unos resultados que permitan definir grupos de una forma tan clara como en el caso de la aplicación del análisis factorial a los datos del marco de la encuesta de población activa, por lo que no se reproducen los resultados.

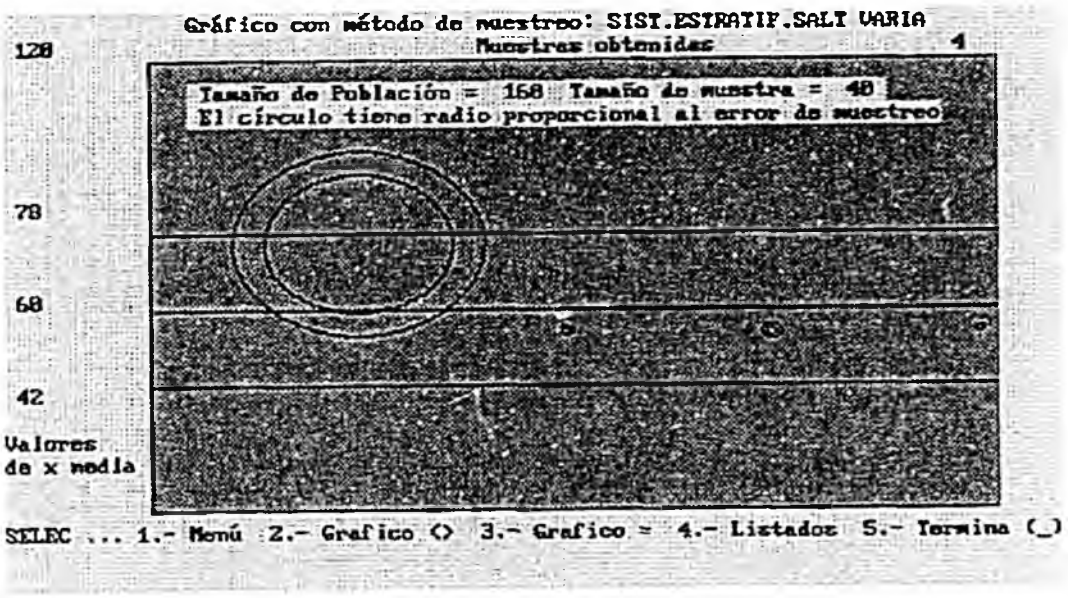
En conclusión, para estos datos, y simplemente observando las cifras y los gráficos, el método que presenta un óptimo en la selección de la muestra es el muestreo sistemático estratificado con intervalo variable, que permite pasar de un error calculado con las reiteraciones (Er.re) de 19,01 a 2,17, y de una estimación del error basado en datos muestrales (Er.m) del 17,20 al 8,25

13.4. DETECCIÓN DE UNIDADES AUTOREPRESENTADAS

Hemos establecido al presentar el capítulo que en el caso de las encuestas económicas tiene especial importancia el que los estratos estén definidos correctamente. Es muy frecuente el tener que aislar ciertas unidades que, por su peso en el conjunto de la población , distorsionan mucho los resultados si no son tratadas convenientemente. Lo más recomendable es formar un estrato con unidades autorepresentadas, unidades con probabilidad uno de pertenecer a la muestra, y analizar ese estrato de manera exhaustiva.

Si por algún motivo dentro de un estrato, que consideramos en principio homogéneo, se ha introducido una unidad que presenta una característica muy diferenciada del resto de las unidades, el programa POSDEM puede utilizarse para detectar estas unidades "anómalas". En el fichero que hemos utilizado con datos del marco de la encuesta industrial las 160 empresas tenían un tamaño medido en número de empleados entre 20 y 170 personas. Vamos a suponer que

la unidad 160 presenta un tamaño de 1000 trabajadores, esto tendrá como consecuencia que en el gráfico donde se muestran las reiteraciones una de ellas, la reiteración que contiene a esta unidad presente un error de muestreo muy elevado

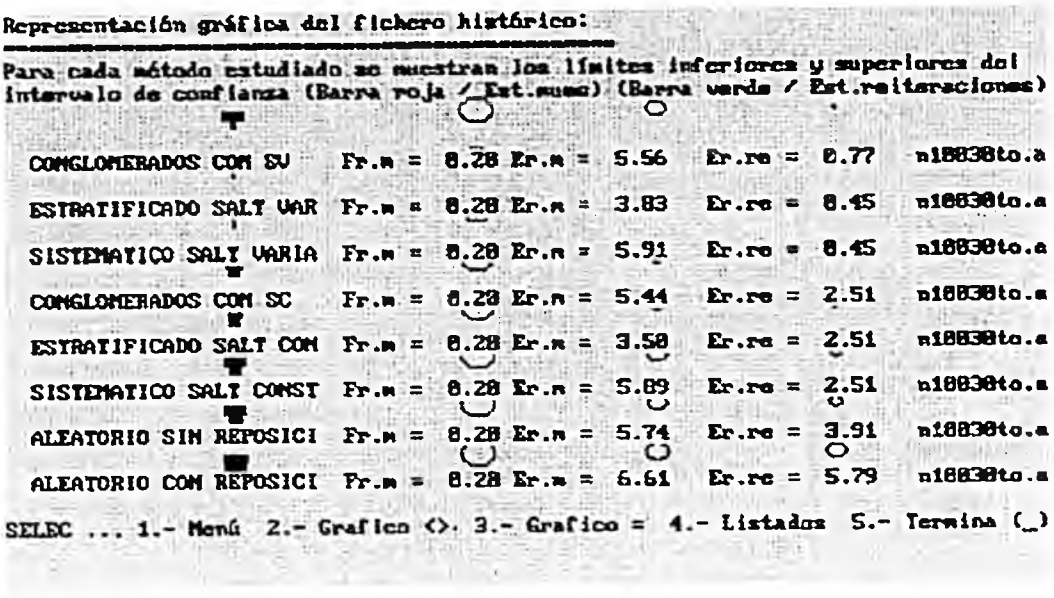


Siempre que al utilizar la aplicación POSDEM se observe un hecho similar, esto es que en el gráfico uno o más valores aparezcan desproporcionados, esto debe hacer que la persona encargada del diseño de la encuesta recapacite sobre los datos que maneja y estudie nuevamente la estratificación original. En nuestro caso

investigaríamos esta unidad de 1000 trabajadores como parte de un estrato independiente de unidades autorepresentadas. Sin necesidad de exagerar, como ha sido el caso, una distribución "anormal" en el gráfico que presenta los principales resultados del experimento deben llevar a reconsiderar la estructura del marco.

13.5. EJEMPLOS DE APLICACIONES A POBLACIONES MARCO CON DISTRIBUCIÓN NORMAL

En la aplicación se han incluido unos ficheros que contienen datos de poblaciones normales, ordenados de menor a mayor, en unos casos y con orden aleatorio en otro. Estos ficheros se han denominado N-media-Desviacion-o-a.asc. En este caso vamos a representar los resultados para el fichero N10030o.asc, esto es una población normal de media 100, desviación 30 y datos ordenados



En este caso, de una población que se distribuye normal de parámetros media 100 desviación típica igual a 30 el error basado en reiteraciones (Er.re) ha pasado de 5,79 en el muestreo aleatorio con reposición a 0,45 en el muestreo sistemático con intervalo variable y la estimación del error basado en datos muestrales (Er.m) de 6,61 a 3,83 para los mismos métodos.

13.6. EJEMPLOS DE APLICACIONES A POBLACIONES MARCO CON DISTRIBUCIÓN BINOMIAL

La aplicación POSDEM incorpora la posibilidad de trabajar con datos cualitativos. Así, se han incorporado ejemplos con poblaciones binomiales que toman valores cero o uno según las ausencia o presencia de un cierto carácter en la unidad referida. El nombre de los ficheros sigue las mismas reglas que en el apartado anterior.

Concretamente en la aplicación se han incorporado dos ficheros: BIN1001o.ASC y BIN1005o.ASC donde el último carácter del nombre es o y no cero, quiere decir que los datos están ordenados. En el primer caso se trata de 100 valores de los cuales sólo los cinco primeros toman el valor uno y todos los demás cero, se trata de un distribución binomial de parámetros 100 y proporción 0.05. En el segundo caso se trata de 100 valores donde la mitad toma

el valor uno y los restantes cero. Para ejecutar esta opción es necesario recordar que previamente a seleccionar un fichero de datos reales cualitativos es necesario señalar al programa que se trata de datos cero, uno, con la opción F12.

A continuación se ha realizado un gráfico con los datos relativos a los distintos métodos, sus varianzas calculadas con reiteraciones (VR) , con valores muestrales (VM), y con las misma estimación de varianza (VV). En las cifras se han eliminado cuatro decimales.

Gráfico 20.- Comparación de varianzas para los datos de una población binomial de parámetros 100 y p=0.05. La burbuja es proporcional al error basado en datos muestrales.

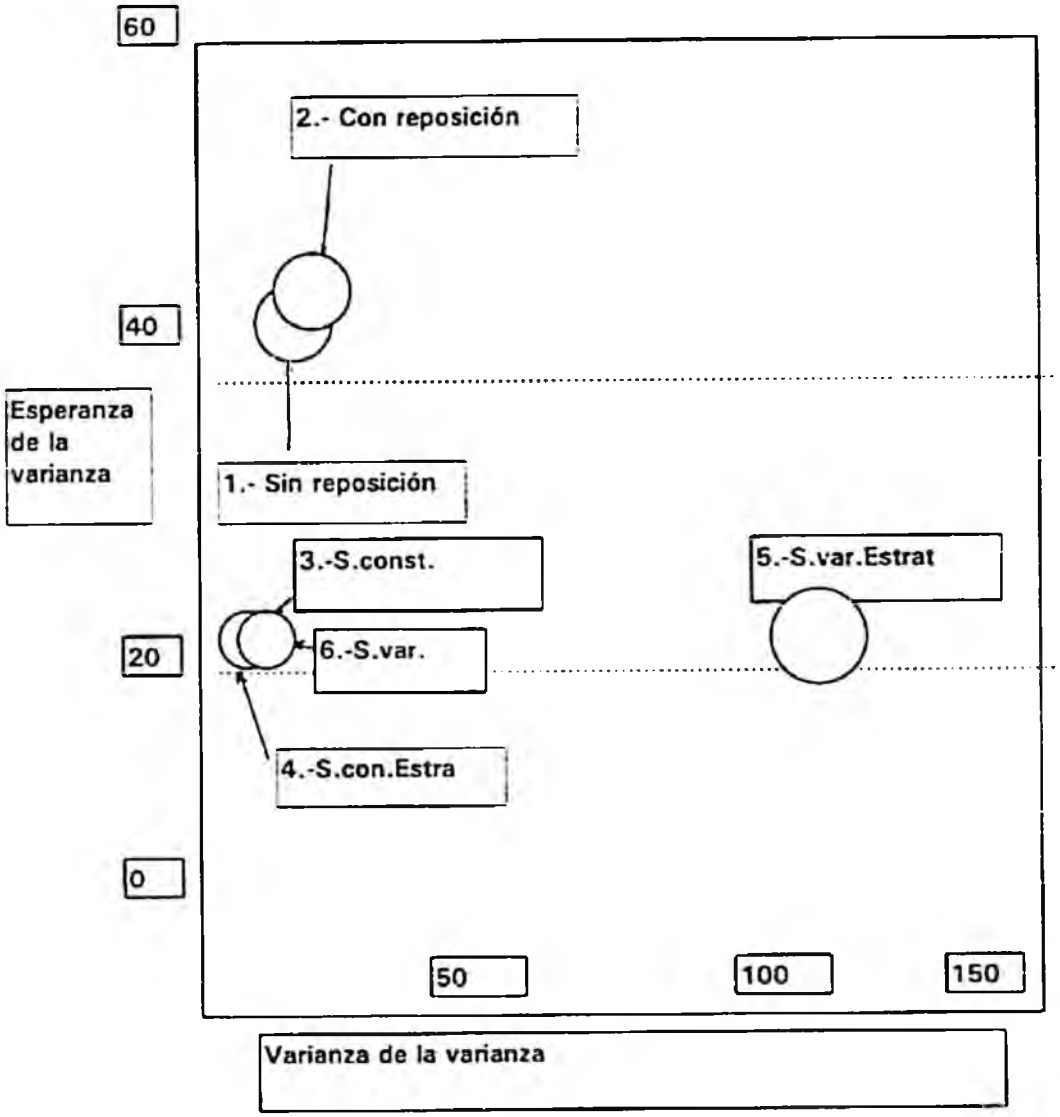


Tabla 34.- Varianzas y estabilidad para una población binomial.

Método	Vmed	Vmedr	Vvr
1.-SR	54	41	23
2.-CR	52	44	31
3.-SC	40	25	16
4.-SCE	40	25	20
5.-SVE	110	25	127
6.-SV	40	25	16

Los resultados son suficientemente explícitos. La interpretación de estos resultados es difícil y puede estar en relación con la población natural que se esta utilizando como marco.

CAPÍTULO 14

MUESTREO BIETÁPICO

14.1. MUESTREO POLIETÁPICO

En este capítulo se va a estudiar el siguiente problema: en una primera etapa se obtiene una muestra por el procedimiento de selección con probabilidades proporcionales al tamaño y, una vez seleccionadas las unidades de primera etapa, se vuelve a realizar un submuestreo para conseguir las unidades últimas, esta vez con un procedimiento de probabilidades iguales. En concreto en la primera etapa se utilizará un tamaño de muestra igual a dos, que es el caso más recomendable cuando el proceso de estratificación se desarrolla al máximo. El esquema de selección en esta primera etapa seguirá los pasos descritos en el procedimiento con reposición parcial y probabilidades proporcionales al tamaño de Sánchez-Crespo y Gabeiras que se ha demostrado en los capítulos anteriores más eficaz que los otros métodos con los que se ha comparado.

Vamos a seguir el procedimiento con un ejemplo:

14.1.1. PRÁCTICA DE MUESTREO BIETÁPICO PARA LA ENCUESTA DE POBLACIÓN ACTIVA

En el caso del diseño de la encuesta de población activa que hemos seguido en el capítulo correspondiente se realizaba una selección de unidades muestrales en el interior de una sección . La sección se define como un área geográfica perfectamente delimitada para la que se dispone de un listado de viviendas en el que figura el número de personas que habitan cada vivienda. En el diseño de una encuesta, normalmente esta sección en la cual se selecciona la muestra es, a su vez, parte de otra muestra. Tanto la primera muestra de secciones, muestra de unidades de primera etapa, como la segunda submuestra de viviendas, muestra de unidades últimas o unidades elementales, contribuyen a la variación total de estimador y por tanto contribuyen a la formación del error de muestreo.

Situándonos en un determinado estrato, donde, por ejemplo se han agrupado las secciones de una provincia en base a ciertas características de forma que disponemos de un conjunto de veinte secciones de similar caracterización respecto a las variables que se hayan utilizado para realizar la estratificación. Se trata de veinte secciones no necesariamente próximas geográficamente pero que mantienen una parecida estructura en cuanto a su estructura interna. Por ejemplo todas ellas pueden ser áreas rurales con bajo nivel industrialización. El número de habitantes y el número de viviendas de

14.1. MUESTREO POLIETÁPICO

En este capítulo se va a estudiar el siguiente problema: en una primera etapa se obtiene una muestra por el procedimiento de selección con probabilidades proporcionales al tamaño y, una vez seleccionadas las unidades de primera etapa, se vuelve a realizar un submuestreo para conseguir las unidades últimas, esta vez con un procedimiento de probabilidades iguales. En concreto en la primera etapa se utilizará un tamaño de muestra igual a dos, que es el caso más recomendable cuando el proceso de estratificación se desarrolla al máximo. El esquema de selección en esta primera etapa seguirá los pasos descritos en el procedimiento con reposición parcial y probabilidades proporcionales al tamaño de Sánchez-Crespo y Gabeiras que se ha demostrado en los capítulos anteriores más eficaz que los otros métodos con los que se ha comparado.

Vamos a seguir el procedimiento con un ejemplo:

14.1.1. PRÁCTICA DE MUESTREO BIETÁPICO PARA LA ENCUESTA DE POBLACIÓN ACTIVA

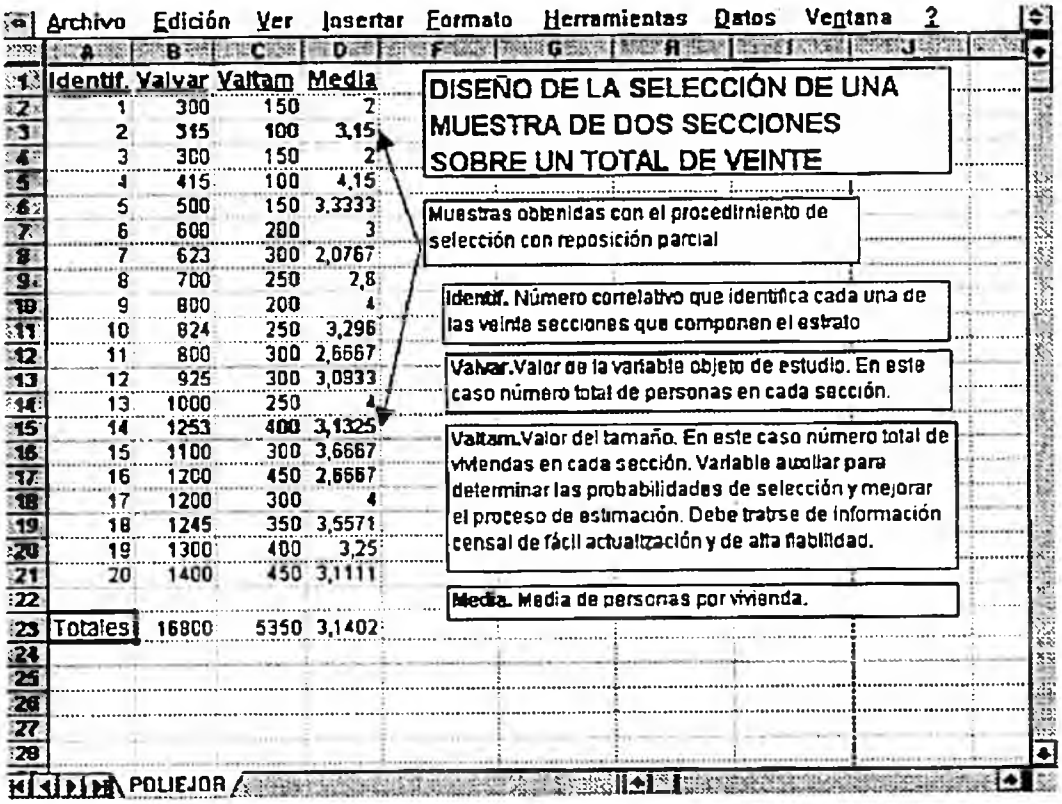
En el caso del diseño de la encuesta de población activa que hemos seguido en el capítulo correspondiente se realizaba una selección de unidades muestrales en el interior de una sección . La sección se define como un área geográfica perfectamente delimitada para la que se dispone de un listado de viviendas en el que figura el número de personas que habitan cada vivienda. En el diseño de una encuesta, normalmente esta sección en la cual se selecciona la muestra es, a su vez, parte de otra muestra. Tanto la primera muestra de secciones, muestra de unidades de primera etapa, como la segunda submuestra de viviendas, muestra de unidades últimas o unidades elementales, contribuyen a la variación total de estimador y por tanto contribuyen a la formación del error de muestreo.

Situándonos en un determinado estrato, donde, por ejemplo se han agrupado las secciones de una provincia en base a ciertas características de forma que disponemos de un conjunto de veinte secciones de similar caracterización respecto a las variables que se hayan utilizado para realizar la estratificación. Se trata de veinte secciones no necesariamente próximas geográficamente pero que mantienen una parecida estructura en cuanto a su estructura interna. Por ejemplo todas ellas pueden ser áreas rurales con bajo nivel industrialización. El número de habitantes y el número de viviendas de

cada sección es la información que se va a utilizar como marco para seleccionar la muestra.

En el cuadro que muestra la figura podemos observar el conjunto de las veinte secciones que forman las unidades de primera etapa en el estrato, identificadas por un número correlativo. Junto a estos valores figuran las variables que se van a emplear para realizar la selección de la muestra. Toda esta información es censal. Si se desea seguir este ejemplo con al aplicación POSDEM el fichero que contiene estos datos es POLIEJOR:ASC.

Tabla 35.- Muestreo bietápico: selección de las unidades de primera etapa.

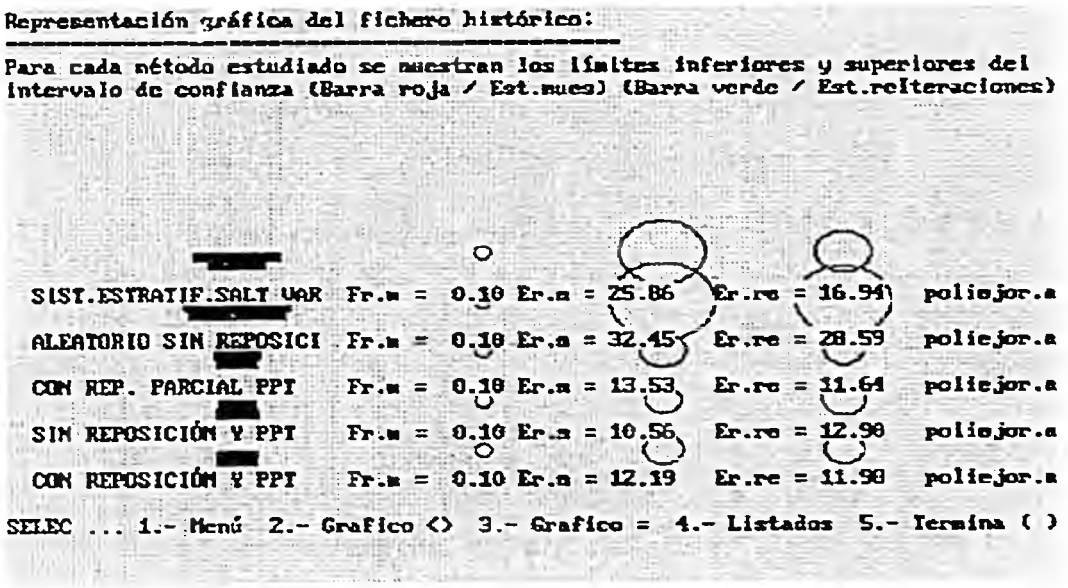


	Identif.	Valvar	Valtam	Media
1	1	300	150	2
2	2	315	100	3,15
3	3	300	150	2
4	4	415	100	4,15
5	5	500	150	3,3333
6	6	600	200	3
7	7	623	300	2,0767
8	8	700	250	2,8
9	9	800	200	4
10	10	824	250	3,296
11	11	800	300	2,6667
12	12	925	300	3,0833
13	13	1000	250	4
14	14	1253	400	3,1325
15	15	1100	300	3,6667
16	16	1200	450	2,6667
17	17	1200	300	4
18	18	1245	350	3,5571
19	19	1300	400	3,25
20	20	1400	450	3,1111
21				
22	Totales	16800	5350	3,1402

14.2. MÉTODO DE SELECCIÓN DE LAS UNIDADES DE PRIMERA ETAPA: ESQUEMA DE SÁNCHEZ-CRESPO Y GABEIRAS

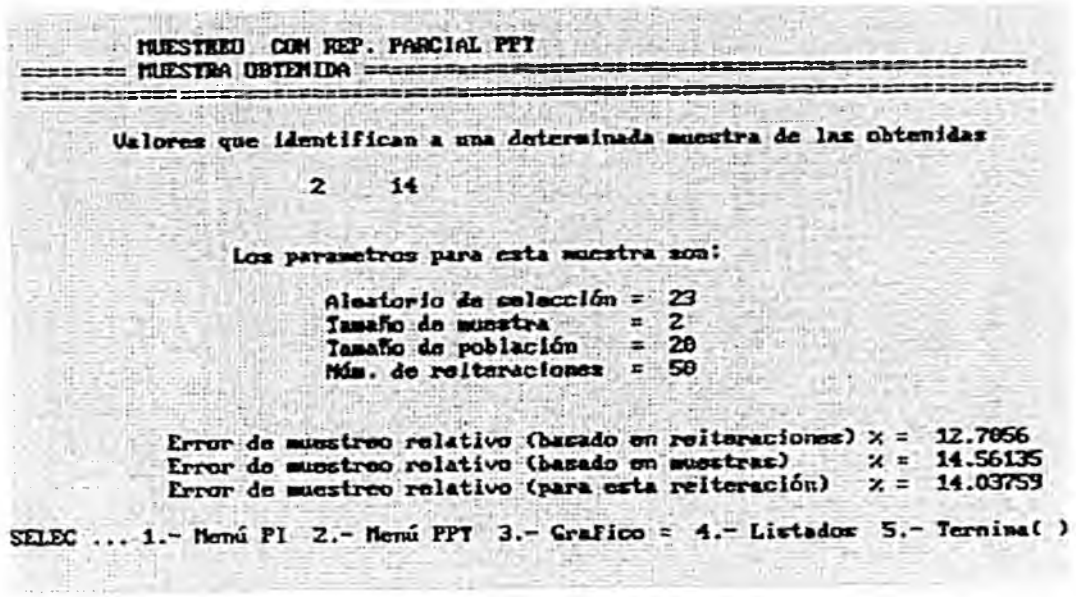
Si suponemos que el número de unidades de primera etapa está fijado en dos unidades para cada estrato de veinte unidades, podemos con la aplicación POSDEM hacer un estudio comparativo de métodos para determinar en base a la estructura de población que estamos investigando cuál será el método que mejor se ajusta.

Con los pasos descritos en el capítulo sobre el esquema con reposición parcial es posible llegar a obtener la siguiente representación gráfica del fichero histórico.



Leyendo el fichero histórico de la ultima línea a la primera, vemos que los tres primeros métodos estudiados, los métodos de probabilidades desiguales, presentan un error que varía entre el 11% y el 14%. A continuación y sobre los mismos datos se ha estudiado el método aleatorio sin reposición que presenta un error bastante más elevado que cercano al 29% . Por último se ha aplicado a los mismos datos el método sistemático con estratificación y intervalo variable que presenta un error del 17% Es necesario resaltar que para este último método el caso de muestras de tamaño dos es el caso más desfavorable. Por tanto, con el tamaño de muestra fijado de dos unidades y esa estructura poblacional conviene aplicar un método de probabilidades desiguales. Por las consideraciones realizadas en el apartado sobre indicadores de superpoblación elegimos el método de Sánchez-Crespo y Gabeiras, con reposición parcial y probabilidades proporcionales al tamaño.

Con el procedimiento descrito en el apartado correspondiente al esquema de Sánchez-Crespo y Gabeiras, de esta población se selecciona una muestra de tamaño dos. La pantalla donde se seleccionan una muestra concreta se reproduce a continuación y es resultado de pulsar simultáneamente las teclas ctrl+F.



El estimador para el total es

$$\hat{x}_{scg} = \frac{1}{2} \cdot \sum_{i=1}^{2} \frac{\hat{x}_i}{p_i}$$

Y la varianza estimada

$$\hat{V}(\hat{x}_{scg}) = B + W$$

donde

$$B = \frac{M - nb}{M} \cdot \frac{1}{n(n-1)} \cdot \left[\sum_{i=1}^{n} \left(\frac{\hat{x}_i}{p_i} - \hat{x}_{scg} \right)^2 \right]$$

$$W = \frac{nb}{M} \left[\frac{1}{n^2} \cdot \sum_{i}^{n} \frac{\hat{v}_2(\hat{x}_i)}{p_i^2} \right]$$

$$\hat{V}(\hat{x}_i) = \frac{M-2}{M} \cdot \frac{s_i^2}{2}$$

14.3. MÉTODO DE SELECCIÓN DE LAS UNIDADES ÚLTIMAS: ESQUEMA SISTEMÁTICO CON INTERVALO VARIABLE.

Una vez seleccionadas las dos secciones que van a formar la muestra de unidades de primera etapa, disponemos de un listado actualizado de las viviendas de cada sección, donde figura el número de personas por hogar; en la misma línea de lo expuesto cuando se trató el ejemplo de aplicación a los datos de la encuesta de población activa, este listado procede del cuaderno de tabulación del agente censal. El nombre de los ficheros que contienen estos datos son respectivamente EPA100.ASC y EPA400.ASC para la sección muestral número dos, que tiene en su interior cien viviendas y para la sección muestral número catorce, que esta compuesta por cuatrocientas viviendas.

Siguiendo el ejemplo, que se ha seguido paso a paso en el capítulo de la aplicación a la encuesta de población activa, con la primera sección, y una vez obtenida la muestra, con la opción CTRL+P, se puede entrar a la gestión del muestreo bietápico prevista en la aplicación.

```
Muestra obtenida para:  EL MUESTREO  SIST.ESTRATIF.SALI VARIA

Valores que identifican a una determinada muestra de las obtenidas

          1              1              2              3
          3              4              4              4
          5              5

                Los parametros para esta muestra son:

Aleatorio de selección =   2              Tamaño de muestra      =   10
Tamaño de población    =  100             Núm. de reiteraciones  =   10

        Error de muestreo relativo (basado en reiteraciones) % =  3.549315
        Error de muestreo relativo (basado en muestras)      % =  9.642122
        Error de muestreo relativo (Para la reiteración)     % =  6.629126

 Estinación de la media =  3.2   Varianza de la media   =  .045

SELEC ... 1.- Gestión muestreo polietápico 2.- Volver barra inferior      (_)
```

Si se selecciona la opción uno, del menú de barra inferior, pasaremos a una nueva pantalla que permite distintas opciones relacionadas con la muestra bietápica.

```
                        Muestreo polietápico

          Menú para la gestión del fichero de unidades primarias:

          1.- Listar ficheros con resultados de las unidades primarias
          2.- Borrar fichero de unidades primarias
          3.- Calcular el estimador con reposición parcial
          4.- Volver opción barra inferior
          5.- Termina

                                        SELEC ...  (_)
```

Si se repite el proceso para la segunda sección, se puede obtener un listado que representará las dos unidades muestrales, en las que a su vez se ha realizado un submuestreo.

```
Estimadores del total y varianzas para cada unidad primaria

Valores que caracterizan a las dos muestras obtenidas

                Los parametros para estas muestras son:

Fichero de datos para la 1ª unidad ..... epa100.asc

                Estimador del total para la 1ª unidad    =     310.000000
                Varianza del estimador para la 1ª unidad =     225.000000

Fichero de datos para la 2ª unidad ..... epa400.asc

                Estimador del total para la 2ª unidad    =    1260.000000
                Varianza del estimador para la 2ª unidad =    5700.000000

SELEC..Volver Menú: 1.- gestión del muestreo polietápico 2.- inicial PUSDEM  ( )
```

Una vez que se dispone de las estimaciones de cada sección y de su error de muestreo podemos agrupar toda la información con el procedimiento de cálculo que se genera al seleccionar la opción tres del menú de gestión del proceso polietápico..

```
        12          925          300          5.607476E-02
        13          1000         250          4.672897E-02
        14          1253         400          7.476635E-02
        15          1100         300          5.607476E-02
        16          1200         450          8.411215E-02
        17          1200         300          5.607476E-02
        18          1245         350          6.542056E-02
        19          1300         400          7.476635E-02
        20          1400         450          8.411215E-02

¿ Identificación de la primera unidad muestral ? 2
¿ Identificación de la segunda unidad muestral ? 14

                Valores      Valores      Prob.      Estimadores
 Identif.     de la var.   del tamaño    (Mi/M)     Total     Varianza
        2       315.00       100.00       0.02     310.00       225.00
       14      1253.00       400.00       0.07    1260.00      5700.00

 Estimador del total para el esquema SCG .....                16718.75
 Desviación del estimador del total para el esquema SCG ....    181.02
 Límite superior de confianza para el 95% .....              17080.79
 Límite inferior de confianza para el 95% .....              16356.71
SELEC..Volver Menú: 1.- gestión del muestreo politápico 2.- inicial POSDEM (_)
```

Por tanto se obtiene , para la variable número de personas , una estimación del total con su correspondiente error de muestreo. El límite inferior será 16.356 y el superior será 17.080 para una confianza del 95% En el caso de que en una nueva realización en el tiempo que tuviese otro resultado en más o menos un 3% sobre el valor del estimador, se podría establecer que esta variación es significativa. Por ejemplo, si al mes de haber realizado esta encuesta se plantea su repetición para hacer un seguimiento de la serie, y en esta nueva realización el resultado arroja una caída del 5 %, esto es un estimador del total de 15.882 , en el caso de que el error de muestreo fuese similar, tendríamos un intervalo de confianza del 15.522 a 16.242 personas Como el valor anterior cae fuera del intervalo podríamos asegurar con una confianza del 95 % que el cambio no ha sido debido al proceso de variabilidad introducido al trabajar con muestras. Esto es muy

frecuente en encuestas sobre el desempleo. Hay que resaltar que en este ejemplo hemos trabajado con una muestra que investigaba cuarenta viviendas y que ha permitido obtener resultados bastante precisos para un colectivo de casi diecisiete mil personas.

14.3.1. PRÁCTICA DEL MUESTREO BIETÁPICO PARA LA ENCUESTA INDUSTRIAL

Consideramos el siguiente caso: Se trata de un estrato formado a su vez por conglomerados. La formación del estrato y de los conglomerados se realiza de manera que las empresas incluidas en el estrato son lo más homogéneas posible entre sí y los conglomerados se han formado de tal forma que las empresas que pertenecen a un determinado conglomerado son heterogéneas entre sí. Los datos del problema son los siguientes:

El número total de empresas del estrato es de 560. Dentro del estrato se han formado un total de 4 conglomerados. Dentro de cada conglomerado hay en media 140 empresas. El método de selección de la muestra es un esquema bietápico. En la primera etapa el tamaño de muestra es igual a dos conglomerados por estrato y el método de selección es con reposición parcial y probabilidades proporcionales al tamaño. En la segunda etapa se obtendrá dentro de cada conglomerado una muestra sistemática estratificada con intervalo variable y probabilidades iguales, con una fracción de muestreo igual al 25%.

Esta información censal se ha incorporado a un fichero para su tratamiento informático , el nombre del fichero es EIEPPT.ASC , sus valores y los principales resultados del estudio, pueden observarse en la siguiente pantalla.

Los parametros para estas unidades son:

Identificación	Valores de la variable	Valores del tamaño	Probabilidades (Mi/M)
1	4872	120	.2142857
2	5400	130	.2321429
3	5900	150	.2678571
4	6206	160	.2857143

¿ Identificación de la primera unidad muestral ? 1
¿ Identificación de la segunda unidad muestral ? 4

Identif.	Valores de la var.	Valores del tamaño	Prob. (Mi/M)	Estimadores Total	Varianza
1	4872.00	120.00	0.21	4876.00	1890.00
4	6206.00	160.00	0.29	6180.00	3000.00

Estimador del total para el esquema SCG 22192.33
Desviación del estimador del total para el esquema SCG 434.79
Límite superior de confianza para el 95% 23061.92
Límite inferior de confianza para el 95% 21322.75
SELEC..Volver Menú: 1.- gestión del muestreo polietápico 2.- Inicial POSDEM ( )

Como paso previo es necesario determinar las dos secciones que van a formar la muestra de unidades de primera etapa, en este caso el resultado ha sido la muestra (1,4), a esto se accede desde el menú de probabilidades proporcionales, una vez seleccionados los datos y el método, con la opción ctrl+F. Después se seleccionan las muestras dentro de los conglomerados. Concretamente estos datos están disponibles en los ficheros EIP160.ASC y Eip120.ASC que contienen la información correspondiente a cada conglomerado.

Se recuerda que los pasos a seguir para gestionar la muestra bietápica son :

- En primer lugar con la opción Ctrl+P gestionar el fichero de experimentos anteriores con muestreo bietápico. Borrar con la opción del menú de gestión de muestras bietápicas.

- Después , desde el menú inicial , con F10 elegir los tamaños de población (160), de muestra(32) y reiteraciones (5), seleccionar la opción de sistemático con la técnica del lazo con intervalo variable y , con F4, definir al programa dónde se encuentran los datos que forman el marco, el fichero EIE160.asc.
- Seleccionar una muestra aleatoria concreta del espacio muestral, esto se realiza con la opción Ctrl+P, que además permite con un menú de barra inferior pasar a un menú de gestión de la muestra bietápica. Se recomienda en este punto obtener un listado de los resultados para la primera muestra.
- Se repiten los pasos del segundo punto pero esta vez con las especificaciones de población (40) y tamaño de muestra (10) del fichero EIE040.ASC.
- Después seleccionar la segunda muestra concreta, y pasar, con ctrl+P, a listar los resultados del experimento. Con la opción cálculo el programa proporciona la pantalla que se ha incluido en el texto con los principales resultados.
- El comentario que se puede hacer de estos resultados es similar al que se realizó en este mismo capítulo al analizar la aplicación a la encuesta de población activa.

CAPÍTULO 15

MUESTREO SISTEMÁTICO CENTRADO CON INTERVALO VARIABLE: CAMBIOS EN LA ESPECIFICACIÓN DE LOS MODELOS

15.1. INTRODUCCIÓN

Presentamos en este capítulo un nuevo método de selección de unidades muestrales que, en términos de error cuadrático medio, mejora al resto de los métodos sistemáticos con los que se ha comparado, en los estudios empíricos que hemos llevado a cabo. Resaltaremos que mejora el comportamiento del método centrado propuesto por Madow para k par y el sistema de correcciones debido a Yates. El programa de ordenador, "POSDEM" se utiliza como instrumento de verificación. Este programa lo hemos realizado, entre otros fines, para optimizar el proceso de diseño de encuestas por muestreo probabilístico en poblaciones finitas. Y se utilizará, en este capítulo, para determinar, bajo el enfoque de los modelos de superpoblación, que método es preferible a otros, para diferentes especificaciones de una determinada estructura de población y diferentes tamaños de muestra. También se analiza la representatividad de las medidas utilizadas.

Siguiendo a Bellhouse y Rao (1975:694-697) hemos llevado a cabo un análisis del comportamiento de diferentes métodos de selección de unidades muestrales ante cambios en las especificaciones de los modelos. La evaluación se ha realizado sobre diferentes procedimientos sistemáticos de selección en el caso de modelos de superpoblación polinómicos de grado entre uno y cinco,

calculando los valores esperados del error cuadrático medio del estimador sobre un conjunto de realizaciones aleatorias.

Nuestro foco de atención se ha orientado al comportamiento errático observado en el método centrado cuando aumenta el tamaño de muestra y a la cuestión de no tener que diferenciar entre valores pares o impares del número de grupos, $k=N/n$, que se forman en la población. Hasta el momento sólo el método de Yates eliminaba la tendencia lineal para valores pares o impares del tamaño de muestra o del número de grupos en la población. El método centrado de Madow no elimina la tendencia cuando k es par. Hemos comprobado, también, que cambios en la especificación del término de error pueden suponer que los métodos centrados sean inestables frente a los restantes métodos considerados.

La población marco esta formada por 128 unidades, caracterizados y ordenados según la población del censo de 1951, la variable de estudio ha sido la población del censo de 1961[1]. Esta población se ha considerado como modelo para ajustar diferentes modelos polinómicas de grado entre uno y cinco. Nuestra intención es comprobar como afectan al error cuadrático medio los cambios en la especificación del modelo o los cambios en el tamaño de muestra.

Los métodos analizados han sido: *aleatorio estratificado con una unidad por estrato, sistemático con intervalo de muestreo constante, sistemático corregido en los extremos de Yates* [2] , *sistemático equilibrado* [3], *sistemático modificado* [4] *y sistemático*

[1] Referenciada en MURTHY, P.R.KRISHNAIAH Y C.R. RAO, 1988.

centrado de Madow[5]. A estos métodos clásicos se han incorporado dos novedades: el primero es un método que hemos definido como intervalo de muestreo variable y, el segundo, y verdadero objeto de esta comunicación, que consiste en aplicar el anterior al método centrado propuesto por Madow, a este método lo hemos denominado: *sistemático centrado con intervalo variable*[6].

Dejamos para otro momento la comparación entre el método centrado con intervalo variable y los métodos con probabilidades proporcionales al tamaño. Adelantamos que es posible realizarla, en los mismo términos que la actual, utilizando la aplicación POSDEM. También es posible analizar el comportamiento de los métodos considerados en presencia de poblaciones con variaciones cíclicas, diferencias en los términos de error y presencia de heterocedásticidad de la perturbación aleatoria. Dejamos fuera del análisis el comportamiento en presencia de variaciones cíclicas a pesar de intuir que el método con intervalo variable, por su propia definición, puede presentar ventajas sobre los restantes métodos.

En la primera tabla de resultados se presentan los resultados del error cuadrático medio esperado sobre el modelo de superpoblación para diferentes tamaños de muestra y diferentes grados de ajuste polinómico. Estos mismos datos, para facilitar su lectura y las comparaciones, se han agrupado por tamaños de muestra en la segunda tabla de resultados. Se presenta una colección de gráficos con objeto de mejorar la comprensión de los datos contenidos en estas tablas. Los resultados son coincidentes con los obtenidos por Bellhouse en cuanto al valor esperado. Hemos calculado

de forma empírica estos valores para diferentes métodos y distintas estructuras poblacionales. Una novedad estriba en incorporando al análisis la posibilidad de cálculo de la representatividad del error cuadrático medio esperado, mediante su varianza respecto del modelo de superpoblación. Los datos correspondientes a este último análisis se presentan en la tercera tabla de resultados.

15.2. UN COMENTARIO SOBRE LAS TABLAS DE RESULTADOS

Nos interesamos por el comportamiento de los diferentes métodos, no para cada caso aislado, sino para la situación en la cual las especificaciones, que hacemos sobre el grado del polinomio o el tamaño de muestra, puedan no ser estrictamente correctas. En la medida que, al apartarnos de una determinada especificación, el método sea más inestable nos encontraremos con métodos menos robustos. Así, por ejemplo, en la primera tabla de resultados podemos observar que para un tamaño de muestra igual a dos, si mantenemos el supuesto de una población con tendencia lineal, el método (3) presenta un error cuadrático medio aceptable. Sin embargo este método presenta un inconveniente en cuanto a cambios de especificación y, así, si la población en realidad estuviese definida por una función polinómica de grado dos, el error aumentaría considerablemente, convirtiendo este método en muy ineficaz en comparación con los otros métodos considerados. Ocurre igual ante aumentos en el tamaño de muestra, en especial para muestras de tamaño impar. En el resto de la tabla, y en especial en los gráficos 21 a 26, podemos observar, en términos generales, como (6) es mejor que (2) y que (5) para cualquier tamaño de muestra y para cualquier supuesto sobre la forma de la población.

Para muestras pequeñas entre dos y cuatro unidades el método (2), para polinomios de grado superior a dos, es más ineficaz que el método (5), ver segunda tabla de resultados. El método (6), en este tramo, se comporta igual o ligeramente mejor que el (5). A partir de muestras de tamaño ocho se puede observar como se invierte esta tendencia y el método (5) es más ineficaz que el (2) independientemente del grado de polinomio. Volvemos a comprobar que el método (6) se ajusta e incluso mejora ligeramente al método mejor, que en este caso es (2). En resumen para muestras pequeñas (5) es mejor que (2) y esta tendencia se invierte para muestras grandes. Por otra parte, independientemente del tamaño de muestra el método (6) se ajusta al mejor de los dos. Los casos más desfavorables consistirían en elegir un método de selección (2) con tamaños de muestra pequeños cuando la población se ajusta a una función polinómica de grado dos o superior. Otra situación desfavorable consistirá en optar por el método (5) cuando el tamaño de muestra sea grande. El caso más favorable, en términos generales, será elegir el método (6) independientemente del grado del polinomio y del tamaño de muestra.

Como ejemplo de utilización de la primera tabla de resultados, gráfico 27 y 28, podemos imaginar una encuesta piloto por muestreo, de forma que no conocemos con certeza el tamaño de muestra que el entrevistador podrá finalmente abarcar, no obstante podemos pensar que la muestra se encontrará entre cuatro y dieciséis unidades. Con los métodos clásicos tendremos que optar por métodos diferentes en función de si el tamaño de muestra es uno u otro. La principal mejora del método (6) es que supera el problema

planteado por la elección entre (5) y (2). Ahora tanto en un caso como en otro tomaríamos el método (6).

Por último observamos, en la tercera tabla de resultados (gráfico 29), que a medida que la dispersión debida al error aleatorio aumenta en el modelo, se produce un aumento, en diferente medida, de la variabilidad de los métodos considerados. Esta pérdida de representatividad del error cuadrático medio esperado puede llevar a conclusiones erróneas si esta información no se incorpora en la evaluación de los diseños alternativos. Para esto, cuando se observa una dispersión acusada en torno a la tendencia proponemos evaluar los métodos con el indicador esperanza respecto del modelo del error cuadrático medio mas dos veces la desviación típica respecto del modelo del error cuadrático medio. Hemos planteado, como ejemplo , en la tercera tabla de resultados, distintos supuestos en cuanto al error aleatorio del modelo. En principio hemos considerado un error distribuido normal de media cero y desviación típica 200 por que se ajustaba bien a los datos observados en la población. No obstante ahora podemos simular distintos escenarios donde este supuesto no se mantiene constante.

15.3. MUESTREO SISTEMÁTICO CENTRADO CON INTERVALO VARIABLE

Este método limita el espacio muestral, cuando k es par, a las dos muestras centrales con probabilidad un medio cada una, y en el caso k impar a las tres centrales, con probabilidad un tercio. Esta es la primera diferencia con (5), la segunda, y más importante es que la selección se realiza aplicando un algoritmo de intervalo variable, que mantiene las probabilidades de cada unidad y que proporciona al método una mayor coherencia en cuanto a precisión y estabilidad. Para valores de $k=2$ y $k=3$ este método coincide con lo que vamos a definir como muestreo sistemático con intervalo variable. Introducimos el método con intervalo variable de la siguiente forma: seleccionado un número aleatorio, i, entre 1 y k, las (n-1) unidades restantes se seleccionan con la siguiente regla:

$$z = i + (j-1)(k+1) - c\,k$$

z = Valor de la unidad muestral seleccionada con intervalo variable

Donde, por definición, los valores que toma c vienen dados por las siguientes situaciones:

- Si $z = jk$ no ha sucedido nunca $c = 0$
- Si $z = jk$ ha ocurrido una vez $c = 1$

- Si $z = jk$ ha ocurrido dos veces $c = 2$ ....

Así, $c = 0,1,2$ ... de acuerdo con el número de veces que ha ocurrido que el valor de z ha sido igual al producto de j por k .

El método centrado con intervalo variable se puede aplicar de la siguiente forma práctica: se forma el espacio muestral completo con el método sistemático clásico con intervalo de muestreo constante. Después, distinguiendo si k es par o impar, se seleccionan dos o tres muestras centrales. Con este espacio muestral reducido se forman las dos o tres muestras posibles, según el valor de k, utilizando esta vez el procedimiento descrito como muestreo sistemático con intervalo variable.

15.4. ESPECIFICACIÓN DE LOS MODELOS

Ahora podemos, mediante el módulo de "simulación de estructuras" de la aplicación POSDEM[2], definir un modelo de superpoblación que explique esta población marco considerada mediante la expresión:

$$X_u = a_0 + a_1 U^1 + a_2 U^2 + a_3 U^3 + a_4 U^4 + a_5 U^5 + e_u$$

donde

U representa las unidades de la población y en este caso toma los valores de 1 a 128.

a_i con $i = 1,2,3,4$ y 5 son los parámetros del modelo calculados inicialmente por mínimos cuadrados, pero que pueden ser definidos con otros procedimientos.

e_u es el término de error aleatorio que en este modelo se ha definido con los valores,

$E_m(e_u) = 0$; $E_m(e_u^2) = \sigma^2$; $E_m(e_u e_v) = 0$ el operador E_m denota la esperanza respecto del modelo.

Este enfoque surge de la necesidad de inferir resultados más allá de lo que representa el análisis de una única población natural, consecuencia de una determinada realización.

[2] Con la utilidad de POSDEM es posible definir modelos no lineales más complejos. Permite utilizar modelos diferentes por tramos de población.

Simulador de estructuras de población en POSDEM. Modelos de superpoblación

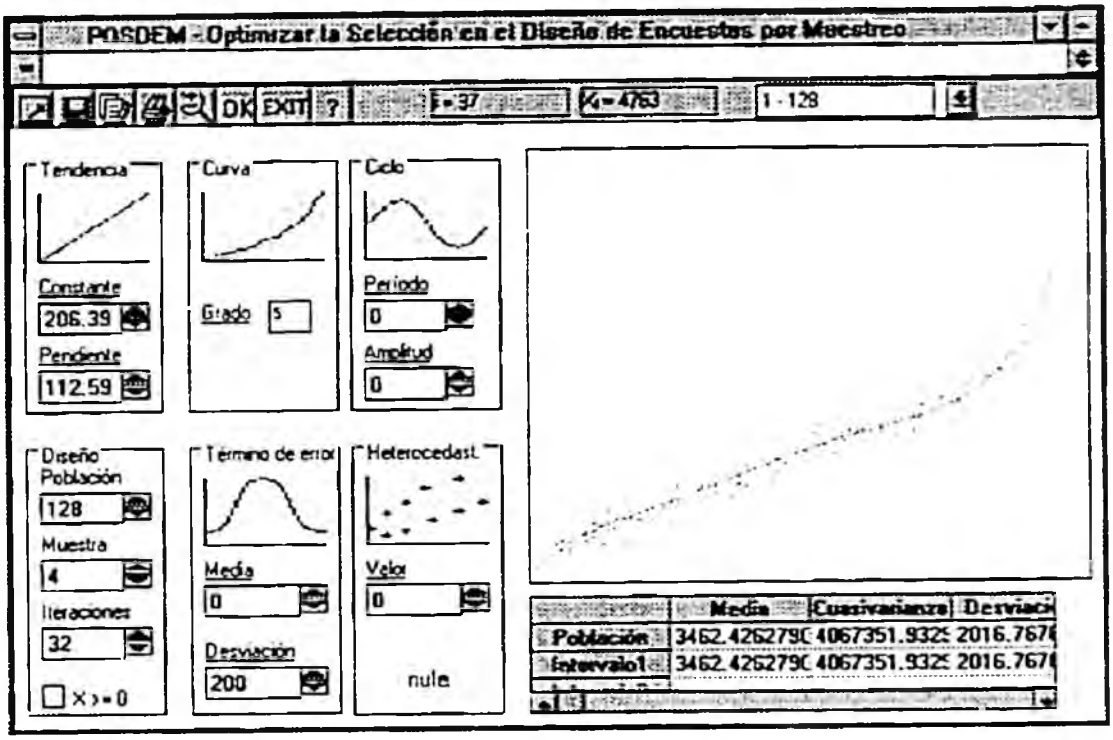


Tabla 36.- Parámetros del modelo para diferentes grados del polinomio.

Modelo	a_0	a_1	a_2	a_3	a_4	a_5
P1	149,08	51,38				
P2	1.179,16	-3,84	0,368			
P3	-83,03	119,02	-1,854	0,0115		
P4	728,48	-2,54	2,347	-0,0390	-0,0002	
P5	206,39	112,59	-3,803	0,0873	-0,0009	-0,000003

Los resultados obtenidos al aplicar estos modelos en la generación de poblaciones aleatorias pueden comprobarse, para el caso lineal, con los siguientes resultados teóricos:

1) Muestreo sistemático:

$$E_m\, V_p\, (\hat{\bar{x}}_{sis}) = a_1^{\,2}\,(k^2 - 1)/12 + \bar{\sigma}^{\,2}$$

con $\bar{\sigma}^{\,2} = \sigma^{\,2}\,(k - 1)/nk$

El primer componente es la varianza debida a la tendencia lineal; el segundo término es la debida al error aleatorio.

2) Muestreo aleatorio:

$$E_m\, V_p\, (\hat{\bar{x}}_{sr}) = a_1^{\,2}\,(k - 1)\,(nk + 1)/12 + \bar{\sigma}^{\,2}$$

3) Muestreo estratificado con una unidad: se asume que la población consiste en *n* estratos formados por los conjuntos de unidades $\{1...k\}$ $\{k+1,...2k\}$,...$\{(n-1)k+1,...,nk\}$. Una muestra aleatoria se toma de cada estrato.

$$E_m\, V_p\, (\hat{\bar{x}}_{str}) = a_1^{\,2}\,(k^2 - 1)/12n + \bar{\sigma}^{\,2}$$

Se puede comprobar que en este caso:

$$E_m\, V_p\, (\hat{\bar{x}}_{strat}) \leq E_m\, V_p\, (\hat{\bar{x}}_{sis}) \leq E_m\, V_p\, (\hat{\bar{x}}_{sr})$$

En el caso de utilizar un modelo parabólico de grado dos, el desarrollo teórico que hemos utilizado para confrontar nuestros resultados ha sido:

$$Em(ecm(2)) - Em(ecm(5)) = (c^2/720)(k^2-1)(19k^2-31) > 0 \text{ para k impar}$$

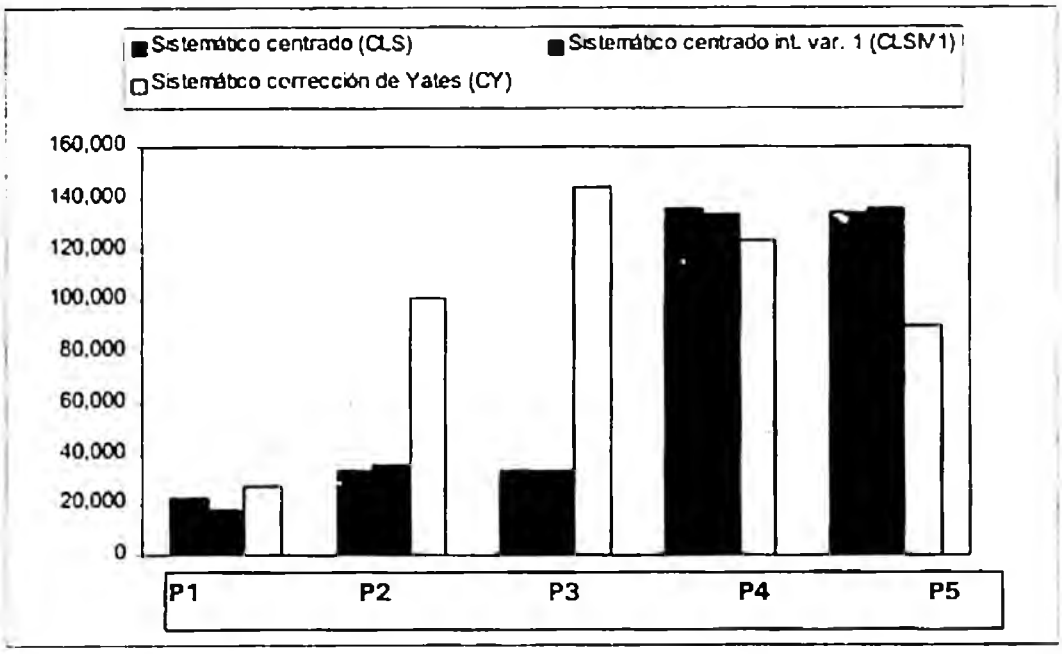
15.5. PRINCIPALES APORTACIONES

Se proporciona un nuevo método que reduce el error cuadrático medio, en términos generales, frente a otros métodos sistemáticos considerados. Esta reducción se lleva a cabo con robustez sobre parámetros que influyen decisivamente en el comportamiento de los otros métodos como son: el tamaño de muestra, el número de grupos en la población y sobre si estos son pares o impares. Se pone de manifiesto la relación acusada entre el término de error aleatorio del modelo y el comportamiento errático del método centrado. Se propone un indicador de la cota superior del error, que incorpora la información relativa a la representatividad del valor esperado, mediante la desviación respecto del modelo del error cuadrático medio.

15.6. ANEXO DE GRÁFICOS[3]

Gráfico 21 a 26.- Esperanza del error cuadrático medio para diferentes grados en la especificación del polinomio, de uno a cinco, y para diferentes tamaños de muestra 2,4,8,16,32.

Gráfico 21.- Para tamaño de muestra n=2: Esperanza del error cuadrático medio para diferentes grados en la especificación del polinomio, de grados uno a cinco.



[3] P1...P5 indicador del grado del polinomio utilizado para especificar el modelo.

Gráfico 22.- Para tamaño de muestra n=4: Esperanza del error cuadrático medio para diferentes grados en la especificación del polinomio, de grados uno a cinco.

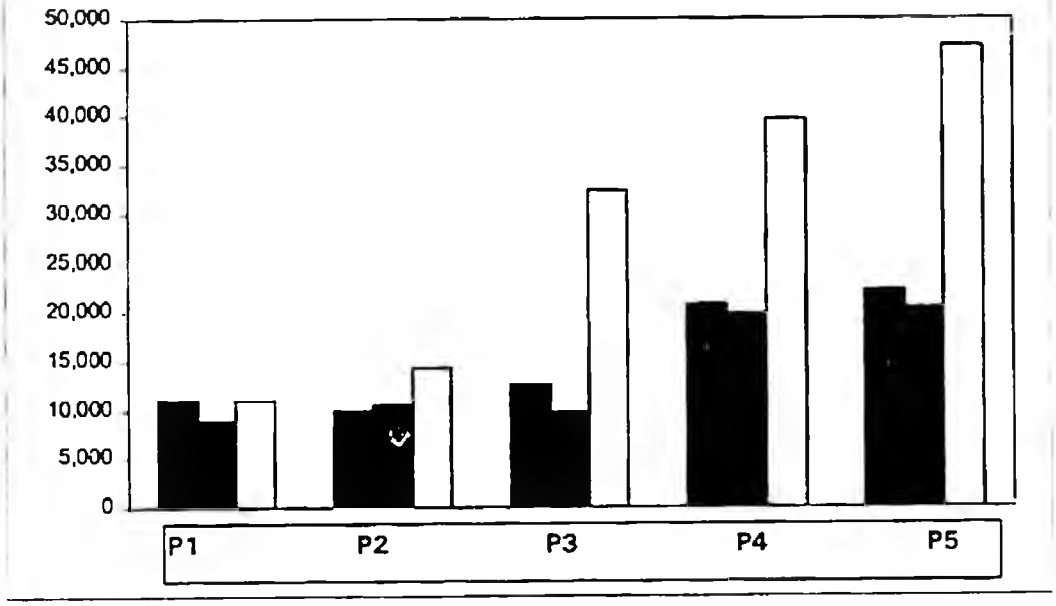


Gráfico 23.- Para tamaño de muestra n=8: Esperanza del error cuadrático medio para diferentes grados en la especificación del polinomio, de grados uno a cinco.

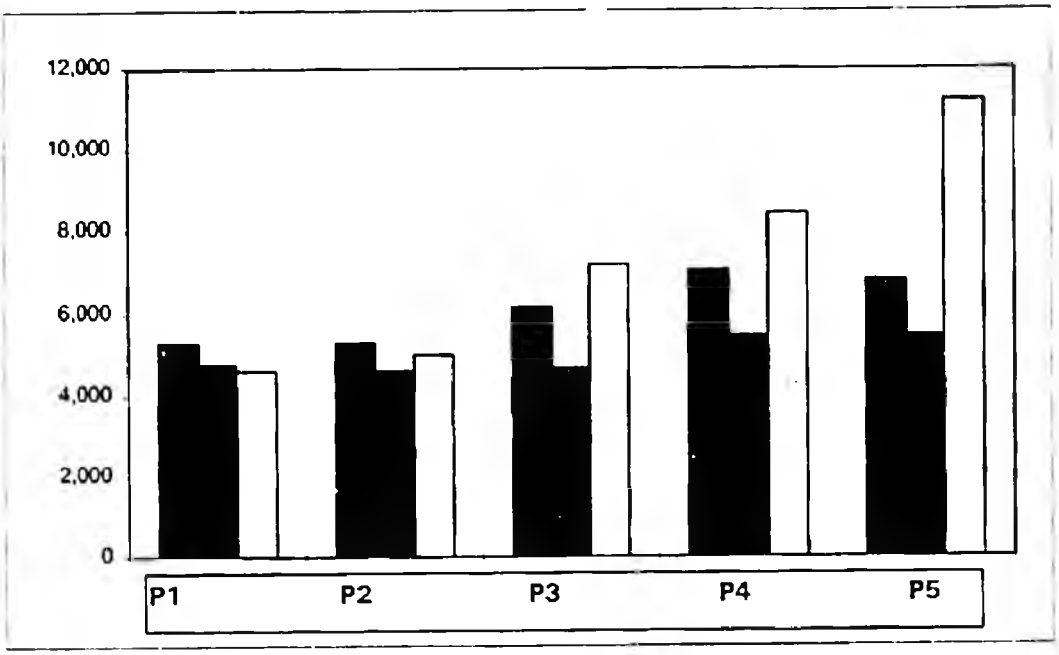


Gráfico 24.- Para tamaño de muestra n = 16: Esperanza del error cuadrático medio para diferentes grados en la especificación del polinomio, de grados uno a cinco.

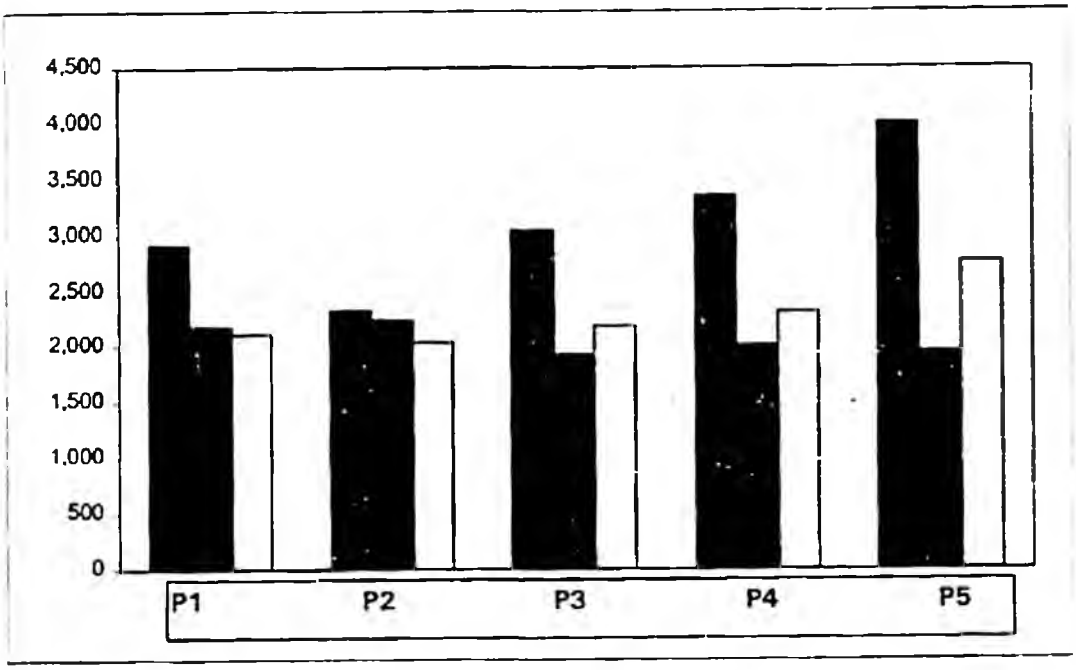


Gráfico 25.- Para tamaño de muestra n=32: Esperanza del error cuadrático medio para diferentes grados en la especificación del polinomio, de grados uno a cinco.

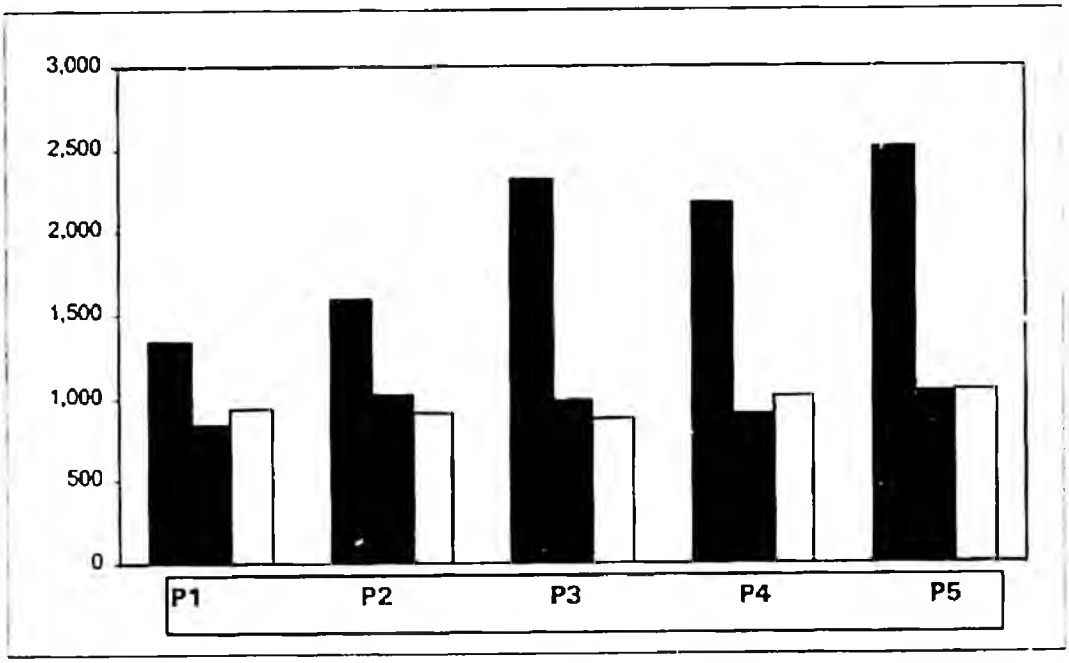
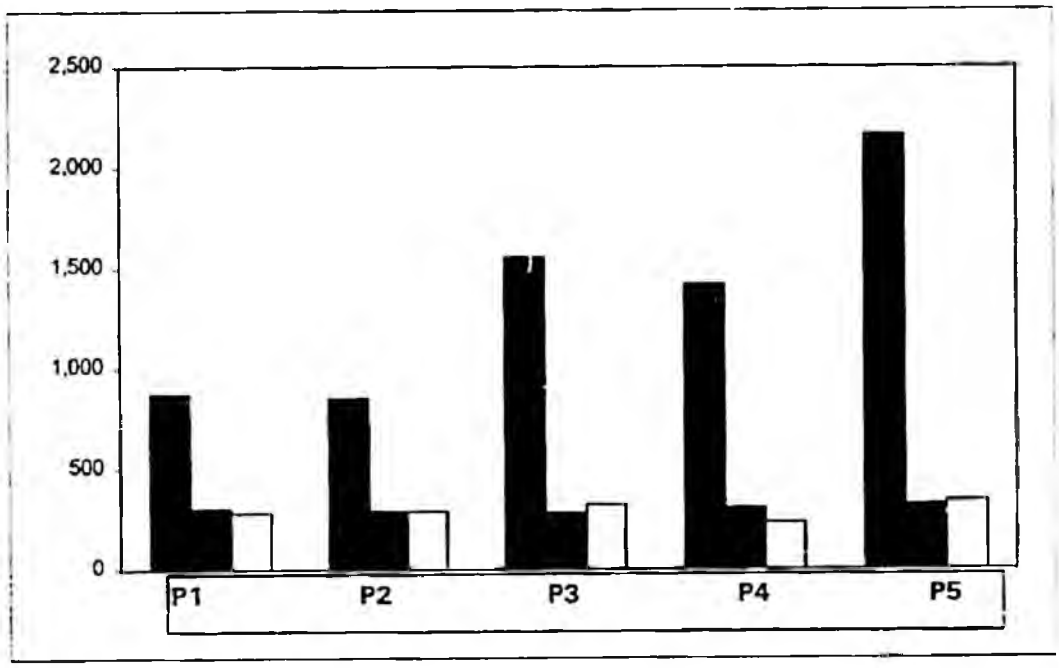


Gráfico 26.- Para tamaño de muestra n=64: Esperanza del error cuadrático medio para diferentes grados en la especificación del polinomio, de grados uno a cinco.



15.7. ANEXO DE GRÁFICOS

Gráfico 27 a 28.- Esperanza del error cuadrático medio para un modelo de polinomio de grado cinco, y para diferentes tamaños de muestra 2,4,8,16,32, 64.

Gráfico 27.- Esperanza del error cuadrático medio para un modelo de polinomio de grado cinco, y para diferentes tamaños de muestra 2,4,8.

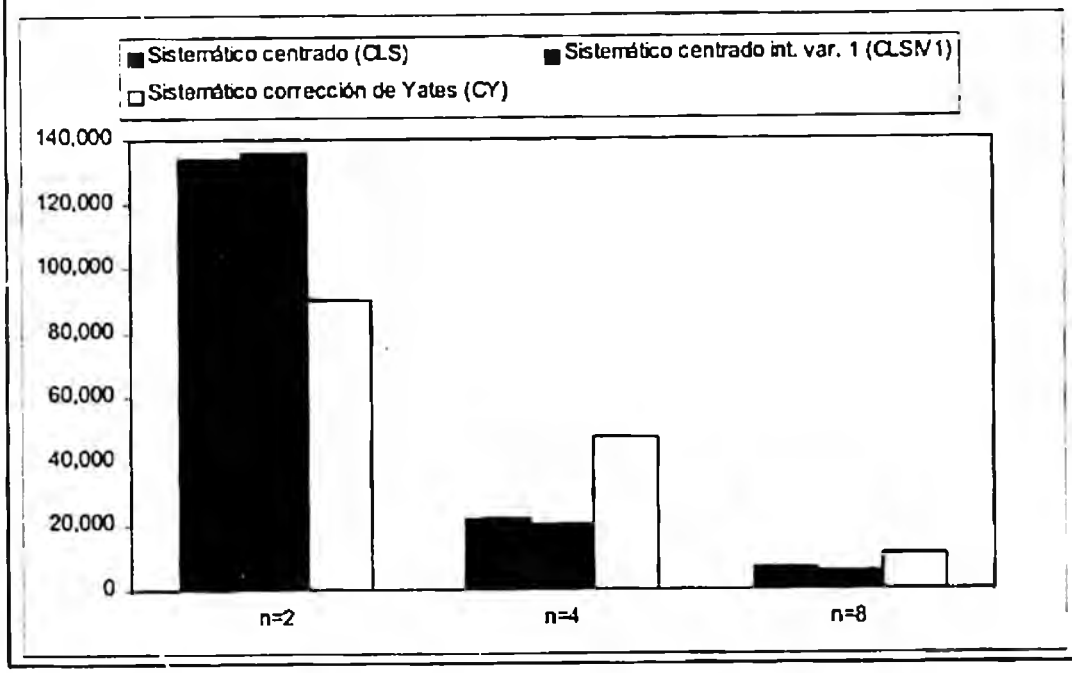
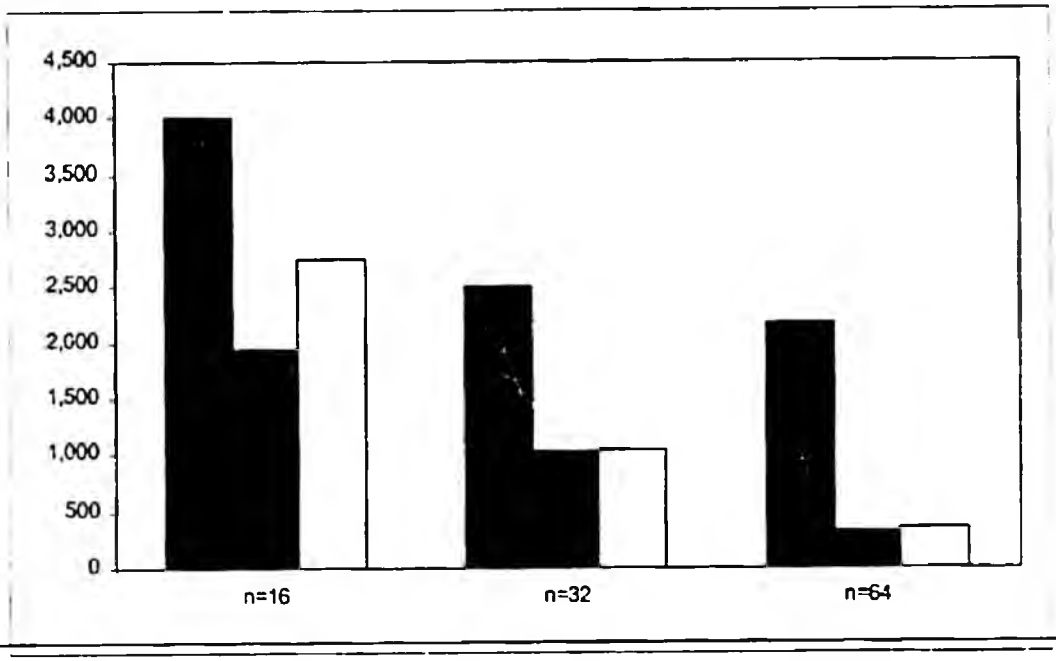
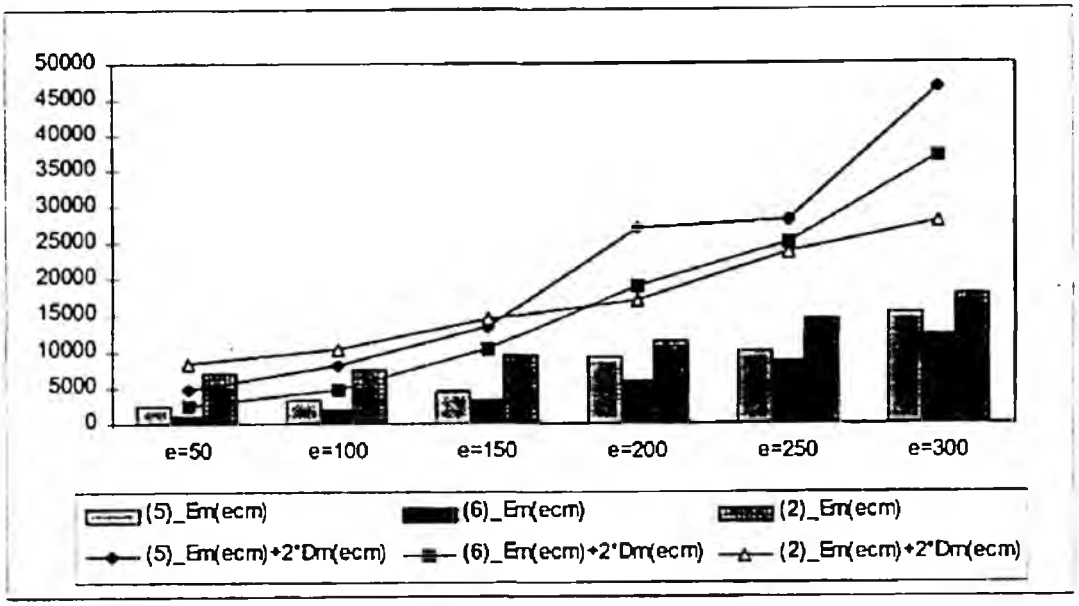


Gráfico 28.- Esperanza del error cuadrático medio para un modelo de polinomio de grado cinco, y para diferentes tamaños de muestra 16,32, 64.



15.8. ANEXO DE GRÁFICOS

Gráfico 29.- Esperanza y varianza del error cuadrático medio respecto de un modelo de polinomio de grado cinco, para tamaño de muestra 8 y diferentes especificaciones de la desviación del error aleatorio (e= 50, 150, 200, 250 y 300)[4]



[4] Em(ecm): Esperanza respecto del modelo del error cuadrático medio
Dm((ecm): Desviación típica respecto del modelo del error cuadrático medio.

15.9. PRIMERA TABLA DE RESULTADOS

Tabla 37.- Esperanza respecto de distintos modelos del error cuadrático medio para diferentes tamaños de muestra.

P 1	n=2	n=4	n=8	n=16	n=32	n=64
Muestreo sin reposición	1,807,470	889,392	430,091	200,803	85,930	28,575
Muestreo estratificado con una unidad	468,966	65,925	11,664	3,029	1,051	314
Sistemático intervalo cte.	917,836	234,827	60,567	15,841	4,578	876
Sistemático intervalo var.	877,473	184,199	18,911	2,027	1,093	302
Sistemático centrado (5)	22,316	11,131	5,306	2,914	1,346	876
Sistemático centrado int. var. (6)	17,709	8,909	4,738	2,195	838	302
Sistemático corrección de Yates (2)	27,364	11,086	4,626	2,124	940	283
Sistemático equilibrado	20,012	10,009	4,588	2,151	952	302
Sistemático modificado	20,012	9,801	4,696	2,273	964	367
P 2	n=2	n=4	n=8	n=16	n=32	n=64
Muestreo sin reposición	1,912,055	938,142	454,189	211,761	90,693	30,239
Muestreo estratificado con una unidad	568,945	80,523	13,586	3,284	1,067	329
Sistemático intervalo cte.	928,912	235,137	60,106	15,935	4,418	858
Sistemático intervalo var.	892,343	190,913	24,333	2,387	875	293
Sistemático centrado (5)	33,236	9,828	5,296	2,324	1,599	858
Sistemático centrado int. var. (6)	36,112	10,560	4,631	2,234	1,016	293
Sistemático corrección de Yates (2)	101,332	14,417	5,016	2,043	913	299
Sistemático equilibrado	221,264	21,372	5,338	2,175	993	293
Sistemático modificado	221,264	56,899	16,772	5,117	1,711	465

P 3	n=2	n=4	n=8	n=16	n=32	n=64
Muestreo sin reposición	2,007,522	989,823	478,829	223,067	95,740	31,820
Muestreo estratificado con una unidad	908,524	155,275	24,111	4,512	1,212	349
Sistemático intervalo cte.	1,496,334	417,512	109,196	27,021	6,755	1,559
Sistemático intervalo var.	1,468,792	344,809	32,764	2,316	874	290
Sistemático centrado (5)	33,220	12,658	6,183	3,040	2,322	1,559
Sistemático centrado int. var. (6)	33,567	9,753	4,644	1,931	982	290
Sistemático corrección de Yates (2)	144,150	32,342	7,210	2,193	881	321
Sistemático equilibrado	219,826	22,378	5,533	1,961	953	290
Sistemático modificado	219,826	58,541	17,055	5,264	1,748	468
P 4	n=2	n=4	n=8	n=16	n=32	n=64
Muestreo sin reposición	2,043,795	1,001,458	485,771	226,754	97,323	32,366
Muestreo estratificado con una unidad	945,669	184,749	29,631	5,379	1,315	349
Sistemático intervalo cte.	1,590,946	423,138	109,281	27,456	7,311	1,417
Sistemático intervalo var.	1,560,935	351,663	39,008	4,321	984	304
Sistemático centrado (5)	136,220	20,771	7,034	3,350	2,173	1,417
Sistemático centrado int. var. (6)	133,559	19,783	5,456	2,007	901	304
Sistemático corrección de Yates (2)	123,425	39,740	8,477	2,309	1,003	238
Sistemático equilibrado	285,637	114,280	13,567	2,670	916	304
Sistemático modificado	285,637	88,042	24,273	6,983	2,011	783
P 5	n=2	n=4	n=8	n=16	n=32	n=64
Muestreo sin reposición	2,053,174	1,005,942	489349	227754	98146	32667
Muestreo estratificado con una unidad	907,932	206,361	36,596	6,364	1,459	363
Sistemático intervalo cte.	1,497,070	474,571	129,125	32,926	8,615	2,165
Sistemático intervalo var.	1,460,596	409,723	47,769	5,224	1,000	322
Sistemático centrado (5)	134,145	22,177	6,781	4,005	2,514	2,165
Sistemático centrado int. var. (6)	135,632	20,430	5,432	1,949	1,030	322
Sistemático corrección de Yates (2)	90,126	47,207	11,213	2,758	1,042	349
Sistemático equilibrado	283,291	111,368	13,480	2,588	971	322
Sistemático modificado	283,291	87,113	24,573	6,868	2,232	587

15.10. SEGUNDA TABLA DE RESULTADOS

Tabla 38.- Esperanza respecto de distintos modelos del error cuadrático medio para diferentes tamaños de muestra.

n=2					
	P 1	P 2	P 3	P 4	P 5
Sistemático centrado (5)	22,316	33,236	33,220	136,220	134,145
Sistemático centrado int. var. (6)	17,709	36,112	33,567	133,559	135,632
Sistemático corrección de Yates (2)	27,364	101,332	144,150	123,425	90,126
Sistemático equilibrado	20,012	221,264	219,826	285,637	283,291
Sistemático modificado	20,012	221,264	219,826	285,637	283,291
n=4					
	P 1	P 2	P 3	P 4	P 5
Sistemático centrado (5)	11,131	9,828	12,658	20,771	22,177
Sistemático centrado int. var. (6)	8,909	10,560	9,753	19,783	20,430
Sistemático corrección de Yates (2)	11,086	14,417	32,342	39,740	47,207
Sistemático equilibrado	10,009	21,372	22,378	114,280	111,368
Sistemático modificado	9,801	56,899	58,541	88,042	87,113
n=8					
	P 1	P 2	P 3	P 4	P 5
Sistemático centrado (5)	5,306	5,296	6,183	7,034	6,781
Sistemático centrado int. var. (6)	4,738	4,631	4,644	5,456	5,432
Sistemático corrección de Yates (2)	4,626	5,016	7,210	8,477	11,213
Sistemático equilibrado	4,588	5,338	5,533	13,567	13,480
Sistemático modificado	4,696	16,772	17,055	24,273	24,573

n = 16					
	P 1	P 2	P 3	P 4	P 5
Sistemático centrado (5)	2,914	2,324	3,040	3,350	4,005
Sistemático centrado int. var. (6)	2,195	2,234	1,931	2,007	1,949
Sistemático corrección de Yates (2)	2,124	2,043	2,193	2,309	2,758
Sistemático equilibrado	2,151	2,175	1,961	2,670	2,588
Sistemático modificado	2,273	5,117	5,264	6,983	6,868
n = 32					
	P 1	P 2	P 3	P 4	P 5
Sistemático centrado (5)	1,346	1,599	2,322	2,173	2,514
Sistemático centrado int. var. (6)	838	1,016	982	901	1,030
Sistemático corrección de Yates (2)	940	913	881	1,003	1,042
Sistemático equilibrado	952	993	953	916	971
Sistemático modificado	964	1,711	1,748	2,011	2,232
n = 64					
	P 1	P 2	P 3	P 4	P 5
Sistemático centrado (5)	876	858	1,559	1,417	2,165
Sistemático centrado int. var. (6)	302	293	290	304	322
Sistemático corrección de Yates (2)	283	299	321	238	349
Sistemático equilibrado	302	293	290	304	322
Sistemático modificado	367	465	468	783	587

15.11. TERCERA TABLA DE RESULTADOS

Tabla 39.- Esperanza y varianza del error cuadrático medio respecto de un modelo de polinomio de grado cinco, para tamaño de muestra 8 y diferentes especificaciones de la desviación del error aleatorio (e= 50, 150, 200, 250 y 300)[5]

e=50	Em(ecm)	Vm(ecm)	Em(ecm)+2*Dm(ecm)
Sistemático centrado (CLS)	2412	1308214	4699
Sistemático centrado int.var. (CLSIV1)	1146	537681	2613
Sistemático corrección de Yates (CY)	6857	469300	8228
e=100	**Em(ecm)**	**Vm(ecm)**	**Em(ecm)+2*Dm(ecm)**
Sistemático centrado (CLS)	3407	5551375	8119
Sistemático centrado int.var. (CLSIV1)	1946	2006160	4779
Sistemático corrección de Yates (CY)	7606	1619986	10151
e=150	**Em(ecm)**	**Vm(ecm)**	**Em(ecm)+2*Dm(ecm)**
Sistemático centrado (CLS)	4312	20595906	13388
Sistemático centrado int.var. (CLSIV1)	3423	12149275	10394
Sistemático corrección de Yates (CY)	9447	6247733	14446

[5] Em(ecm): Esperanza respecto del modelo del error cuadrático medio
Dm((ecm): Desviación típica respecto del modelo del error cuadrático medio.

e = 200	Em(ecm)	Vm(ecm)	Em(ecm) + 2*Dm(ecm)
Sistemático centrado (CLS)	9175	79173321	26971
Sistemático centrado int.var. (CLSIV1)	5744	42558181	18791
Sistemático corrección de Yates (CY)	11251	8196990	16977
e = 250	Em(ecm)	Vm(ecm)	Em(ecm) + 2*Dm(ecm)
Sistemático centrado (CLS)	10069	82019657	28182
Sistemático centrado int.var. (CLSIV1)	8669	65641842	24873
Sistemático corrección de Yates (CY)	14443	20746505	23553
e = 300	Em(ecm)	Vm(ecm)	Em(ecm) + 2*Dm(ecm)
Sistemático centrado (CLS)	15405	245961297	46771
Sistemático centrado int.var. (CLSIV1)	12193	154397365	37044
Sistemático corrección de Yates (CY)	17650	25928618	27834

BIBLIOGRAFÍA

ABAD DE SERVIN, A.; (1978); *Introducción al muestreo*; Limusa; Mexico DF.

ALBERDI ALONSO, J.; (1969); *Metodologia de investigación por muestreo*; Euroamericana D.L.; Madrid.

ANTHONY Y. C.K.; (1983); "Double Bootstrap Estimation of Variance Under Systematic Sampling with probability proportional to size"; *The Journal of Statistical Computation and Simulation. Volumen 31, number 2.*

AZORIN POCH, F.; (1969); *Curso de muestreo y aplicaciones*; Aguilar. Madrid.

AZORIN, F.; SÁNCHEZ-CRESPO J.L.; (1986); *Métodos y aplicaciones del muestreo*; Alianza Universidad Textos; Madrid.

BARNETT, V.; (1974); *Elements of sampling theory*; The English Universities Press; London.

BELLHOUSE,D.R.; RAO, J.N.K.; (1975); "Systematic sampling in the presence of a trend"; *Biometrika 62, 694-697*

BREWER, K.R.W.; HANIF, M.; (1983); *Sampling with unequal probabilities*. Springer-Verlag.

BUCKLAND, W.R.; (1951); "A review of the literature of systematic sampling". *Journal of the Royal Statistical Society. Series B, Vol 13, p. 208-215.*

BROMAGHIN, J.; MCDONALD, L.; (1992); "Systematic Encounter Sampling. A Simulation study"; *The Journal of Statistical Computation and Simulation. Volumen 40, Number 384.*

CASSEL ET AL.; (1977*); Foundations of inference in survey sampling*; Wiley Series in Probability and Mathematical Statistics; New York.

COCHRAN,W.G.; (1946); "Relative accuracy of systematic and stratified random samples for a certain class of populations"; Ann. Math.Stat., Vol.17, pp.164-177.

COCHRAN, W.G.;(1977); Sampling Techniques; Third Edition; Wiley. New York.

COCHRAN, W.G.;(1986); *Técnicas de muestreo*; Mexico: Continental.

COWDEN, D.J.; (1957); *Statistical Methods in Quality Control*; Prentice-Hall, NJ.

DEMING, W.E.; (1960); *Sample Design in Business Research*; Jhon Wiley and Sons, New York.

DIACONIS P.; **EFRON B.**; (1983); "Métodos intensivos por ordenador"; *Investigación y Ciencia, p 70-83.*

DUNN, R.; **HARRISON, A.R.**; (1993); "Two-dimensional systematic sampling of land use"; *Journal of the Royal Statistical Society. Series C, Vol 42, nr 4, p. 585-601.*

ESCUDER VALLS, R.; (1992); *Manual de Teoría de probabilidades con nociones de muestreo e inferencia;* Valencia: Tirant la blanche.

FERNANDEZ TROCONIZ, A; (1987); *Probabilidades Estadistica Muestreo*; Tebar Flores, Madrid, 471 Páginas, 1ª Edición, España, Isbn 8473600789.

FINNEY, D.J.; (1950) ; "An example of periodic variation in forest sampling". *Forestry, 23, p 96-111.*

FISHMAN, G.; (1996); "Montecarlo, Concepts, Algorithms and Aplications"; Springer Series in Operation Research Ed. P. Glynn.

FOREMAN, E.K.; BREWER, K.R.W.; (1971); *"The efficient use of suplementary information in standard sampling procedures".* Journal of the Royal Statistical Society. B33, 391-400.

GARCIA ESPAÑA, E.; (1974); *Diseño de la encuesta general de población*. Instituto Nacional de Estadística; Madrid.

GRANGER, C.W.J.; SIKLOS, P.L.; (1XXX) "Systematic sampling, temporal aggregation, seasonal adjustment, and cointegration"; *Wilfrid Laurier University. Department of Economics. Working paper. 19p.*

HAJEK, J.; (1981); *Sampling from a finite population*; New York, Clav-Dupec:M.Dekker.

HANNAN, E.J.; (1962); "Systematic sampling"; *Biometrika 49, 281-283.*

HANSEN, M.H. Y HURWITZ, W.N.; (1943); *"On the theory of sampling from a finite population".* Annals of Mathematical Statistics, 14, 33-362.

HANSEN, M.H.; HURWITZ, W.N.; MADOW. W.G.; (1953); *Sample Survey Methods and Theory;* John Wiley and Sons, New York, Vols. I y II.

HARTLEY, H.O.; (1966); "Systematic sampling with unequal probabilities and without replacement"; Journal of the American Statistical Association *61, 739-748.*

HARTLEY, H.O.; RAO, J.N.K.; (1962); "Sampling with unequal probabilities and without replacement"; Annals of Mathematical Statistics *33, 350-374.*

HASEL, A.A.; (1938); "Sampling error in timber surveys"; *Journal of Agricultural Research 57, 713-736.*

HEDAYAT A.S.; SINHA B.K.; (1991) *Design and inference in finite population*; Wiley and Sons;1.

HEILBRON, D.C.; (1978); "Comparison of estimators of the variance of systematic sampling"; *Biometrika 65, 429-433.*

HENDRICKS, W.A.; (1956); *The matematical Theory of Sampling*; Scarecrow Press, New Brunswick, N.J.

HENRY, G. T; (1990); *Practical Sampling*; Sage Publications, Californi, 139 Páginas, 1ª Edición, Estados Unidos, Isbn 0803929587;0

HIDIROGLOU, M.A. ; GRAY, G.B. ; (1975); "A computer algorithm for joint probabilities of selection"; *Survey Methodology (Statistics Canada) 1, 99-108.*

HORVITZ, D.G.; THOMPSON, D. J.; (1952); "A generalization of sampling without replacement from a finite universe"; Journal of the American Statistical Association *47, 663-685.*

HOTELLING; SOLOMONS; (1932); "Limits of a measure of skewness"; Ann.Math.Stat.; Vol3, pp.141-142.

HYMAN, H.H.; (1954); *Interviewing in Social Research*; University of chicago Press, Chicago, III;1

IACHAN, R.; (1980); "Topics in systematic sampling"; *University of california-Berkeley Ph.D.disertation.*

IACHAN, R.; (1980); "The weight scaling problem - Systematic sampling from a population of truckloads of trees"; *Univ. of Wisconsin Tech. Report No. 626.*

IACHAN, R.; (1981 A); "An asymptotic theory of systematic sampling"; *Univ. of Wisconsin Tech. Report No. 632;1.*

IACHAN, R.; (1981 B.); "An asymptotic comparison between systematic sampling and simple random sampling"; *Univ. of Wisconsin Tech. Report No. 635.*

IACHAN, R.; (1982); "Systematic sampling - A critical review"; *ISR 50, 293-303;1.*

INDIAN COUNCIL OF AGRICULTURAL RESEARCH, NEW DELHI; (1950); "Estimation of catch of marine fish in a sample of landing centres"; *Report on the pilot sample survey conducted on the Malabar coast for estimating the catch of marine fish (unpublished).;1.*

ISAKI, C.T.; PINCIARO, S.J.; (1977): "Numerical comparison of some estimators of the variance under pps systematic sampling"; *In: Proceedings of the Social Statistics Section, American Statistical Association. I, 308-313.;1.*

JONES, A.E.; (1948); "Systematic sampling of continuous parameter populations"; *Biometrika 35, 283-290;1.*

JONES, H. L.; (1955); "The application of sampling procedures to business operations"; Journal of the American Statistical Association *50, 763-776;1.*

JOURNEL, A.G.; HUIJBREGTS, C.J.; (1978); "Mining Geostatistics"; *Academic Press, London.*

JOWETT, H.H.; (1952); "The accuracy of systematic sampling from conveyor belts"; *Applied Statistics 1, 50-59.*

KALTON, G.; (1983); *Introduction to survey sampling*; Beverly Hills: Sage Publications, cop.

KISH, L.; (1965); *Survey Sampling*; New York: Wiley.

KISH, L.; (1975); *Muestreo De Encuestas*; Trillas, 1ª Edición; México.

KONIJN, H.S.; (1973); *Statistical Theory of Sample Survey Design and Analysis*; Amsterdam: North Holland.

KONIJN, H.S.; (1973); "Statistical Theory of Sample Survey Design and Analysis"; *North-Holland, Amsterdam.*

KOOP, J.C.; (1971); "On splitting a systematic sample for variance estimation"; Annals of Mathematical Statistics *42, 1084-1087.*

KOOP, J.C.; (1976); "Systematic sampling of two-dimensional surfaces and related problems"; *Research Triangle Institute, NC.*

KRISHNAIAH, P. R.; RAO, C.R.; (1988); *Handbook Of Statistics. Vol. 6: Sampling;* North-Holland, Amsterdam, 594 Páginas, 1ª Edición, Holanda.

LAHIRI, D.B.; (1954A); "On the question of bias in systematic sampling"; *Proceedings of the World Population Conference 6, 349-362.*

LAHIRI, D.B.; (1954B); "Technical report on some aspects of the development of the sample design"; *National Sample Survey Report No. 5, Government of India. Reprinted in Sankhya 14, 264-316.*

LAMBERT, J.M.; (1972); "Theoretical methods for large-scale vegetation survey"; *In: J.N.R. Jeffers, ed., Mathematical Models in Ecology. Oxford, Blackwell, 87-109.*

LEPKOWSKI J.; BOWLES J.; (1988); "Sampling errror software for personal computers"; *The Survey Statistician.*

LEVY, P.S.; (1991); *Sampling of populations: Methods and applications*; New York: John Wiley & Sons, cop.

LININGER, C. A.; (1984) *La encuesta por muestreo: Teoría y práctica*; Mexico. Continental.

MADOW, L.H.; (1946); "Systematic sampling and its relation to other sampling designs"; Journal of the American Statistical Association *41, 204-217.*

MADOW, W. G.; MADOW, L.H.; (1944); "On the theory of systematic sampling"; Annals of Mathematical Statistics *15, 1-24.*

MADOW, W.G.; (1949); "On the theory of systematic sampling II"; Annals of Mathematical Statistics *20, 333-354.*

MADOW, W.G.; (1953); "On the theory of systematic sampling III"; Annals of Mathematical Statistics *24, 101-114.*

MAHALANOBIS, P.C.; (1944); "On large-scale sample surveys"; *Roy. Soc. Phil. Trans. Ser. B 231, 329-451.*

MAHALANOBIS, P.C.; (1946); "Recent experiments in statistical sampling in the Indian Statistical Institute"; *Journal of the Royal Statistical Society Ser. A 109, 325-378, reprinted in Sankhya (1958), 1-68.*

MAISEL P.; (1995); *Sampling;* Wiley and Sons.

MIRAS, J.; (1985); *Elementos de muestreo en poblaciones finitas*; Madrid, INE.

MOKASHI, V. K.; (1954); "Efficiency of systematic sampling in forest sampling"; *JISAS 6, 101-114.*

MURTHY, M.N.; (1967); *Sampling theory and Methods*; Statistical Publishing Society, Calcutta.

MURTHY, M.N.; SETHI, V. K.; (1965); "Self-weighting design at tabulation stage"; *Sankhya B 27, 201-210.*

NAIR, K. R.; BHARGAVA, R.P.; (1951); "Statistical sampling in timber surveys in India"; *Indian Forest Leaflet, No. 153, Forest Research Institute, Dehradun.*

NORDSKOG, A. W.; CRUMP, S.L.; (1948); "Systematic and random sampling for estimating egg production in poultry"; *Biometrics 4, 223-233.*

ODEH, R. E.; (1983); *Attribute sampling plans, tables of test and confidence limits*; New York: M. Dekker, c.

OSBORNE, J. G.; (1942); "Sampling errors in systematic and random surveys of covertype areas"; Journal of the American Statistical Association *37, 256-264.*

PADAM S.; GARG, J. N.; (1979); "On balanced random sampling"; *Sankhya Ser. C 41, 60-68.*

PATTERSON, H. D.; (1954); "The errors of lattice sampling"; *Journal of the Royal Statistical Society Ser. B 16, 140-149.*

QUENOUILLE, M. H.; (1949); "Problems in plane sampling"; Annals of Mathematical Statistics *20, 335-375.*

RAO, J.N.K.; BAYLESS, D.L.; (1969); *"An empirical study of the stabilities of estimators and variances estimators in unequal probability sampling of two units per stratum".* Journal of the American Statistical Association, 64, 540-559.

RIPLEY, B; (1981); *Spatial statistics*;New York: Wiley.

SÁNCHEZ-CRESPO, G; (1983); "Metodologia para la estimación en dominios de estudio pequeños"; *Memoria de Licenciatura; Universidad Autónoma de Madrid.*

SÁNCHEZ-CRESPO, J.L.; (1994); "Esquema de muestreo con reposición parcial"; *Estadística Española, V 36, p 143-160.*

SÁNCHEZ-CRESPO, J.L.; (1997); "A sampling scheme with partial replacement"; *Journal of official statistics, p 327-339.*

SASAKI, T.; (1981); "Multidimensional systematic sampling";*Journal Inform. Process. 4, 79-88*

SCHEAFFER, R. L.; (1987); *Elementos de muestreo*; Mexico D.F.; Iberoamericana.

SEDRANSK, J.; (1969); "Some elementary properties of systematic sampling"; *Skand. Aktuar. 52, 39-47.*

SETHI, V. K.; (1965); "On optimum pairing of units"; *Sankhya Ser. B 27, 315-320.*

SHAH, B.V.; (1981); *SESUDAAN:Standard errors programs for computing of standardized rates from survey data*; Research triangle Park, NC: Research triangle Institute.

SHAUL, J. R. H.; MYBURGH, C. A. L.; (1948); "A sample survey of the African population of Southern Rhodesia"; *Pop. Studies 2, 339-353.*

SHIUE, G. J.; (1966); "Systematic sampling with multiple random starts"; *Forestry Science 6, 142-150.*

SINGH, D.; PADAM S.; (1977); "New systematic sampling"; *Sankhya Ser. C 40, 72-73.*

SINGH, D.; JINDAL, K. K.; GARG, J. N.; (1968); "On modified systematic sampling"; *Biometrika 55, 541-546.*

STEHMAN, S.V.; OVERTON, W. S.; (1994); "Comparison of variance estimators of the Horvitz_Thompson estimator for randomized variable probability systematic sampling"; *Journal of the American Statistical Association, Vol 89, nr. 425, p. 30-43.*

STEPHAN, F. F.; DEMING, W. E.; HANSEN, M. H.; (1940); "The sampling procedure of the 1940 population census"; Journal of the American Statistical Association *35, 615-630.*

STEPHAN, F.; MCCARTHY, P.J.; *(1XXX); Sampling Opinions*;John wiley and Sons, New York.

SUDAKAR, K.; (1978); "A note on circular systematic sampling"; *Sankhya Ser. C 40, 72-73.*

SUKHATME, P. V.; PANSE, V. G.; SASTRY, K. V. R.; (1958); "Sampling techniques of estimating the catch of sea-fish in India"; *Biometrics 14, 78-96.*

SUKHATME, P.V.; SUKHATME, B.V.; (1970); *Sampling Theory of Surveys with Applications.;* FAO Rome, 2nd ed.

THOMPSON S.; (1992); *Adaptative Sampling*; Wiley.

TIN, M.; THEIN, P.; (1977); "Bamboo inventory in Burma"; *BISI 47, 597-600.*

TÖRNQVIST, L.; (1963); "The theory of replicated systematic cluster sampling with random start"; *RISI 31, 11-23.*

TRUEBLOOD, R.M.; CYERT, R.M.; (1957); *Sampling Techniques in Accounting*; Prentice-Hall, Englewood Cliffs, N.J.

TRYFOS P.; (1996); *Sampling. Methods for applied Research*; Wiley and Sons.

U.N. STATISTICAL OFFICE; (1950); *The preparation of sample survey reports*; Stat.Papers Series C, Núm.1

VALLIANT. R.; (1990); "Comparisons of variance estimators in stratified random en sistematic sampling"; *Journal of official statistics. VOL. 6:2. p. 115-131.*

VERMA, V., SCOTT, C.; O'MUIRCHEARTAIGH, C.; (1980); "Sample designs and sampling errors for the World Fertility Survey"; *Journal of the Royal Statistical Society Ser A 143, 431-473.*

VIZMANOS J.R.; (1984); *Análisis matemático y cálculo de probabilidades*. Fareso. Madrid.

WILLIAMS, R. M.; (1956); "Variance of the mean of systematic samples"; *Biometrika 43, 137-148.*

WOLTER, K. & MCCANN, S.; (1977); "Alternative estimators of variance for systematic sampling"; *Proc. Social Stat. Section, American Stat. Association P II 787-797.*

WU, CHIEN-FU; (1980); "Estimation in systematic sampling with suppelementary observations"; *Univ. of Wisconsin Technical Report.*

YATES, F.; (1948); "Systematic sampling"; *Philosophical Transactions of the Royal Society Ser. A 241, 345-347.*

YATES, F.; (1960); *Sampling Methods for Censuses and Surveys, 3rd edition.*;London: Griffin.

YATES, F.; GRUNDY, P. M.; (1953); "Selections without replacement from within strata with probability proportional to size; *Journal of the Royal Statistical Society Ser. B 15, 253-261.*

ZINGER, A.; (1980); "Variance estimation in partially systematic sampling"; Journal of the American Statistical Association *75, 206-211.*