Muestreo - INEE: Trabajo Grupal

Fecha de entrega: Jueves, 25 de Octubre, 2018

Dr. Emilio López

Baroja Manzano José Luis
Chávez De la Peña Adriana Felisa
Hernández Rodríguez César Alberto

Sección 1 PREGUNTAS ABIERTAS

1.- ¿Cuál es el objetivo principal del muestreo, es decir, en qué situaciones se usa o qué pregunta ayuda a responder el muestreo?

El objetivo del muestreo es extraer elementos de un conjunto acerca del cual se quiere conocer algo (i.e. la población objetivo), de manera que a partir de este subconjunto (i.e. la muestra) se puedan inferir sus características.

2.- ¿Cómo podemos relacionar las siguientes ideas en una sola oración: variabilidad, muestreo, obtención y recolección de datos, estimación, inferencia, población, responder preguntas, precisión, términos probabilísticos, control, medición, parte de la estadística? Es decir, haga una oración que contenga todas las palabras y que a la vez no esté diciendo algo equivocado.

El muestreo es la parte de la estadística que aborda la necesidad de tener un buen control en la obtención y recolección de datos que capturen la variabilidad de la población sobre la cual se quiere conocer algo, y cuya medición permita hacer estimaciones que guíen las inferencias que permitan responder, en términos probabilísticos, las preguntas que se tengan respecto de la misma, con la mejor precisión posible.

3.- ¿Qué diferencia tienen los libros tradicionales de muestreo y el libro de Särndal que estamos utilizando?

Principalmente, que el libro de Särndal ofrece un enfoque generalizado (y no particularizado) del muestreo. Es decir, que a diferencia de los textos tradicionales que ofrecen un "menú" de expresiones matemáticas específicamente desarrolladas para funcionar bajo cierto esquema de selección de muestra, el libro de Särndal presenta una perspectiva más unificada, que aborda el problema del muestreo y el tipo de preguntas a las que con él se busca dar respuesta, a partir de una mirada más generalizada de la función diseño y su relación con la definición de probabilidades de inclusión (y todas las posibles dimensiones en que éstas pueden ser ponderadas), que regulen el muestreo en sí mismo. Además de ello, el libro de Särndal facilita el uso de diversos Softwares especializados en muestreo, en tanto que parten del mismo marco conceptual y comparten la misma terminología.

4.- ¿Qué relación hay entre el software de muestreo en general y el Särndal?

Que utilizan los mismos términos (la misma notación) y parten de un mismo enfoque teórico/conceptual.

5.- ¿Cuál es la principal desventaja de un enfoque articularizado del muestreo en la práctica, en la oficina, en la realidad?

Que dado que las expresiones matemáticas son exclusivas del esquema de selección de muestra con que se esté trabajando, requieren necesariamente que éste se cumpla. Es decir, la forma de trabajo está condicionada al grado en que el muestreo realizado "encaje" en un esquema particular, (y esto no siempre es posible). Por el contrario, un enfoque generalizado permite ajustar el trabajo/desarrollo de las mismas expresiones matemáticas,

Sección 1 continúa en la siguiente página...

en función de las características y restricciones del esquema de muestreo logrado.

6.- Comente en sus palabras cuál sería el procedimiento general o esqueleto del proceso que involucra una encuesta. Como si lo estuviera platicando o explicando a un político o a un joven sin contacto previo con el muestreo.

1. Especificar pregunta de investigación
2. Diseñar instrumento con reactivos adecuados para responder la pregunta
3. Especificar la población objetivo, y los factores que se consideren pueden influir en el valor de la pregunta.
4. Construir un diseño muestral bajo las restricciones en 3.
5. Aplicar los instrumentos.
6. Calcular el valor de la respuesta en la muestra.
7. Extrapolar a la población bajo las restricciones que definieron al diseño.

7.- Proporcione 3 ejemplos sobre el uso del muestreo diferente a una encuesta electoral o de opinión pública. Es decir, se necesitan ejemplos en donde no se trate de una encuesta. En donde no se necesite un cuestionario tal cual como ordinariamente se hace en una encuesta de opinión. De preferencia de ejemplos diferentes a los comentados en clase.

1. La aplicación de una evaluación: Supongamos que queremos conocer el nivel de habilidades matemáticas que tienen los estudiantes de cuarto grado de primaria en México. La población objetivo es finita, pero resulta difícil pensar en la posibilidad de llevar a cabo una evaluación censal. Por lo tanto, me conviene elaborar una estrategia de muestreo que me permita obtener la mayor información posible sobre la variabilidad que presenta este conjunto de estudiantes en la habilidad que interesa medir.
2. Imaginemos que el nuevo C.E.O. de Starbucks quiere realizar un sondeo del estado de las instalaciones con que cuenta su cadena a lo largo del continente, para tener una idea general de los estándares de higiene, calidad y estética que la marca está cumpliendo. Evidentemente, resultaría impráctico esperar a que se pudieran revisar todas y cada una de las sucursales de Starbucks en América, por lo que lo más viable sería realizar un diseño muestral que permita obtener esta información, de una forma más rápida y asequible.
3. En general, al hacer investigación. Cuando en Psicología Experimental se trabaja con tareas experimentales diseñadas para evaluar la ejecución de los sujetos observados en un escenario que emula las propiedades del fenómeno cognitivo o conductual que se quiere estudiar, se busca tener más de un ensayo, de manera que pueda capturar la variabilidad en el comportamiento de cada sujeto; en este sentido, cada ensayo representa un elemento de la muestra a partir de la cual yo busco decir algo sobre el comportamiento del sujeto ante cierto tipo de situaciones (una población infinita compuesta por todos los ensayos posibles). Más aún, considerando que el

Sección 1 continúa en la siguiente página...

fin último de las ciencias del comportamiento es poder hablar de leyes del comportamiento que apliquen para todos los "posibles sujetos" que yo pude haber evaluado, el conjunto particular de los sujetos que de hecho observé se considera una muestra del conjunto total sobre el cual espero decir algo (¡la raza humana!)

8.- Es importante definir bien todos los elementos o detalles involucrados dentro de un ejercicio de muestreo de poblaciones finitas ¿Qué relación tiene esto con el ejercicio de inferir?

Reportar de manera detallada los elementos, controles, características y restricciones tomadas en cuenta durante el proceso de muestreo permite dar validez a los métodos de inferencia estadística empleados. (Por ejemplo, para evaluar si se puede, o no, asumir que se trabajó con muestras aleatorias, etc.)

9.- ¿Qué es un marco muestral y para qué me sirve dentro de la teoría de muestreo?

El marco muestral es una aproximación a la estructura de la población objetivo. Funciona como un esquema general que guía extracción de los elementos contenidos en la población.

10.- ¿Por qué es importante tener un marco muestral de buena calidad y actualizado?

Porque el marco muestral es el esquema con el cual estoy intentando capturar la estructura de la población objetivo. Si el marco está incompleto, desactualizado, o contiene información incorrecta, no importará la calidad del diseño muestral empleado, las conclusiones que de él deriven estarán condenadas a no ser válidas para la población que se pretendía estudiar.

11.- ¿En qué casos tengo problemas con mi marco muestral, cuáles son los típicos problemas que pueden presentarse?

Que el marco esté desactualizado, que contenga elementos que no pueden ser incluidos en muestra, que contenga información incorrecta sobre los elementos de la población, o que simplemente no exista un marco muestral de la población objetivo, entre otros.

12.- ¿Una encuesta me sirve para responder preguntas de un individuo en particular. Sí o no? Explique ampliamente.

En sentido estricto, y muy ingenio, sí. Sin embargo, el objetivo de aplicar una encuesta no es fijar la mirada en los datos recabados respecto de cada individuo, sino en la información que todos ellos, en conjunto, pueden brindarnos sobre las características generales de la población de la que provienen. En otras palabras, al aplicar una encuesta lo que interesa siempre es computar valores muestrales que resuman las características de la población.

13.- ¿Todos los errores en una encuesta tienen que ver con muestreo? ¿Sí o no? Explique ampliamente.

No. De hecho, quizá los más importantes son los errores de medición, o bien, las deficiencias de nuestros instrumentos para medir lo que realmente pretenden medir. Si las preguntas en la encuesta están mal diseñadas (p.ej., son tendenciosas o confusas), los resultados de la encuesta serán erróneos sin importar cuán preciso sea el diseño muestral.

14.- Explique de manera simple las ventajas y desventajas de un enfoque de muestreo basado en diseño.

Se dice que el modelo basado en diseño es "objetivo" en tanto que las características del proceso de muestreo son documentadas en detalle, y son tomadas en cuenta al momento de decidir qué métodos usar para hacer las inferencias correspondientes a los estimadores de interés y sus propiedades. La rigurosa documentación de este proceso lógico da validez a los análisis realizados y a las conclusiones que de éstos se deriven.

Por otro lado, conforme aumenta el tamaño de la muestra, la esperanza de los estimadores tienden al valor verdadero de los parámetros poblacionales de interés (insesgamiento). Esto último implica dos grandes desventajas para la aplicación del muestreo basado en diseño: se requiere de un tamaño de muestra tan grande como sea posible, y por tanto, puede resultar costoso (en términos económicos).

15.- Pensando en un enfoque demuestreo basado en modelos, explique ¿por qué es posible tener tamaños de muestra muy pequeños en este approach

Porque el "modelo" al que hace referencia el nombre "basado en modelos", define las características de la "superpoblación" que genera los datos que componen nuestra población objetivo. Por ejemplo, se puede asumir que la población objetivo proviene de una distribución probabilística "oculta" (definida en nuestro modelo), y por tanto, se pueden inferir las características de la población aún con tamaños de muestra muy pequeños (o incluso iguales a 0) ya que éstas pueden inferirse por medio de simulaciones.

16.- Explique ¿cómo es posible que el enfoque basado en diseño pueda utilizar diseños de muestreo (o probabilidades de inclusión) arbitrarias y a la vez no se considera un enfoque subjetivo?

Porque bajo este enfoque, la estructura estocástica implícita en el muestreo se sitúa en la regla empleada para extraer elementos de la población (i.e. la función diseño y todo lo que de ella deriva). Es decir, en tanto que se tiene un amplio control y documentación de las características del diseño muestral, (desde el uso de diseños estratificados, polietápicos o con probabilidades de inclusión desiguales), se puede confiar en (a.k.a. se "validan") los métodos de inferencia empleados.

17.- ¿Qué es el muestreo probabilístico?

Implica que la muestra ha sido extraída de acuerdo a un mecanismo aleatorio, con algunas restricciones. Tienen que cumplirse las siguientes condiciones:

1. Se puede definir a $S = (S_1, S_2..., S_M)$ (el conjunto de las muestras posibles de

Sección 1 continúa en la siguiente página...

acuerdo con el esquema de selección)

2. Se conoce la probabilidad de selección $p(s)$ de cada muestra posible $s \in S$
3. Del esquema de selección $p(s)$ se desprende la probabilidad de selección π_k de cada elemento k contenido en la población, (siendo que $\pi_k \neq 0$)
4. La muestra s es seleccionada mediante un mecanismo aleatorio que asigna a todas las muestras posibles la misma probabilidad de ser seleccionada ($p(s)$ es constante).

18.- Comente por qué no es posible determinar que una muestra es probabilística si sólo se observa la muestra extraída.

Porque observar la muestra no nos dice nada sobre cómo se consiguió, que es el componente clave del muestreo probabilístico. P. ej., si resulta que tenemos una muestra de 15 personas, no sabemos si esas quince fueron elegidas al azar entre los elementos de cierta población, digamos echando un volado antes de decidir si cada individuo en la población formaría parte de la muestra, o si fueron elegidas arbitrariamente utilizando una regla determinista, por ejemplo, decidiendo que la muestra estaría conformada únicamente por las personas de la población cuyo primer apellido comienza con Z.

19.- ¿Qué son las probabilidades de inclusión?

Representa la probabilidad de que cualquier elemento contenido en la población sea incluído en la muestra.

20.- ¿Qué es el diseño de muestreo?

Representa la estructura formal o matemática detrás de la extracción de una muestra particular.

21.- ¿Cuál es la diferencia entre $p(s)$ y π_k?

$p(s)$ (i.e. la función diseño) define la probabilidad de obtener una muestra particular, en tanto que π_k representa la probabilidad de que un elemento cualquiera, contenido en la población, se incluya en la muestra.

22.- ¿Para qué me sirve determinar $p(s)$ y π_k en todo este asunto del muestreo que vemos en el curso, ¿Qué importancia tiene cada uno en la teoría vista?

La función diseño ($p(s)$) define una distribución de probabilidad sobre el conjunto de todas las muestras posibles S, por lo que constituye el núcleo detrás del proceso de selección de una muestra particular. Asimismo, en cada una de estas muestras se encuentran contenidos ciertos elementos k, que, dada la estructura definida por $p(s)$, tienen cierta probabilidad π_k de aparecer en la muestra observada. El control sobre estos indicadores (por ejemplo, asegurarse que nunca ocurra que $\pi_k = 0$), promueve que las estimaciones a obtener a partir de la muestra con la que estamos trabajando sean más precisas, en

función de qué tan bien capture nuestro diseño muestral la variabilidad de los elementos contenidos en la población objetivo.

23.- ¿Es posible (¿y por que?) utilizar técnicas demuestreo que hemos visto conmuestras no probabilísticas?

El muestreo no probabilístico implica que los elementos contenidos en la muestra fueron seleccionados de manera arbitraria (y no así, aleatoria). En estos escenarios, pareciera que no tendría sentido hablar de las probabilidades de inclusión de cada elemento, o bien, pretender que existe una expresión matemática que "aleatoriamente" lleve a la selección de nuestra muestra observada (es decir, no existe $p(s)$)

24.- ¿Qué es un parámetro (en la teoría de muestreo)?

Un parámetro es una característica (no observable) de la población objetivo. El valor de un parámetro se pretende aproximar extrayendo una muestra de dicha población y definiendo una función (conocida como estimador) cuya esperanza, a través de diferentes muestras, coincide con el valor del parámetro de interés.

25.- ¿Un parámetro tiene variabilidad. Sí, no, por qué?

No: en teoría de muestreo, los parámetros poblacionales son características fijas. Los estimadores para dichos parámetros sí tienen variabilidad porque son funciones de muestras aleatorias.

26.- ¿Y la variable de estudio, es una variable aleatoria? ¿Sí, no? ¿Por qué?

No, en el enfoque basado en diseño para poblaciones finitas se asume que los valores de la variable de estudio en cada uno de los elementos que componen la población (y_k), aunque desconocidos, son fijos. La aleatoriedad de las estimaciones a obtener reside en que las muestras a partir de las cuales éstas son calculadas, son de hecho realizaciones de la variable aleatoria S, aproximada por $p(s)$

27.- ¿Un estimador de un parámetro tiene variabilidad? ¿Sí, no? ¿Por qué?

Sí, porque se trata de una función que depende de los elementos en muestra, y como los elementos en muestra son aleatorios, dicha función también es aleatoria.

28.- Explique cómo es eso de que un estimador estima un parámetro. ¿Qué es un estimador? ¿Cómo funciona con "peras y manzanas" ¿Qué quiero de un estimador y cómo me aseguro de que eso que quiero suceda? Explíquelo a un niño preguntón.

"¿Has visto los partidos de fútbol? Hay mucha gente en el estadio, ¿no? ¿Como cuántas personas crees que están en las gradas? Bueno, pues ahora imagina que te pregunto cuántos años tienen todos esos que van al estadio. Para calcularlo con certeza tendrías que preguntarle a cada persona en el estadio, pero si lo haces así te vas a tardar mucho

y ni te a va dar tiempo porque antes de que termines acabará el partido y medio estadio se irá a su casa sin haberte respondido. Lo que puedes hacer es preguntarle a un grupito de personas, digamos los que se sientan en las butacas rojas, que son pocos y te toma poco tiempo: les preguntas, anotas sus edades, y sacas el promedio de esas edades. Después vas con los de la cabecera de las Chivas, les preguntas lo mismo y vuelves a sacar el promedio de sus edades, y así puedes imaginar que repites con muchos grupos pequeños de fanáticos, sacándole el promedio de edad a cada uno. Bueno, pues resulta que si al final del partido calculas el promedio de esos promedios, vas a quedar muy cerca del promedio de todas las personas en el estadio, ¡pero no tuviste que preguntarle a cada una!"

29.- ¿De donde viene la variabilidad en el muestreo bajo el enfoque basado en diseño?

De la variación entre muestras: la muestra seleccionada se interpreta como una realización de entre todas las muestras posibles, y los resultados obtenidos en dicha muestra se comparan contra aquellos que pudieron ser observados en aquéllas muestras posibles pero no observadas.

30.- ¿La variabilidad en el muestreo basado en diseño la puedo controlar o mínimo describir? ¿Para qué me interesa controlarla o describirla? ¿Cómo? ¿Mediante qué? Explique.

Es posible describirla y controlarla. Es importante conocerla y controlarla ya que al aminorarla, sin importar el método de estimación que utilicemos este no cambiará mucho. De esta manera también tendremos un control sobre el sesgo, lo que nos da un estimador más cerca del dato real. La manera de controlarla es que nuestro diseño la pueda hacer lo más pequeña posible.

31.- ¿Cuál es la diferencia entre un estimador y una estimación?

La estimación representa el valor puntual que se computa a partir de una muestra particular, mientras que el estimador representa el conjunto total de las estimaciones que podrían obtenerse a lo largo de las distintas muestras posibles.

32.- ¿Qué es la distribución muestral? ¿Qué me dice? ¿Es fácil obtenerla siempre. Sí, no, por qué? En caso de que no, ¿Qué puedo hacer entonces?

Se puede definir como la distribución probabilística del estimador dependiendo de la muestra aleatoria. Nos indica la variabilidad del estimador en las diferentes muestras. Debido a que no siempre se conoce el parámetro de cada muestra, no siempre se puede obtener. Sin embargo, empleando el teorema central del límite podemos inferir que la media y varianza del estimador se comportarán como en una normal.

33.- ¿Por qué nos importa estimar en todo momento la media y la varianza de un estimador? ¿Cómo se conecta con el concepto de la distribución muestral?

Porque de acuerdo con el Teorema del Límite Central, con independencia de cuál sea el tipo de distribución que se asume para y_k (los elementos contenidos en la población), las

posibles estimaciones que podría obtener a partir de las distintas muestras posibles se distribuyen de acuerdo a una distribución normal con media en $\hat{\theta}$ (o en el caso de estimadores insesgados, como la media, en θ).

34.- ¿Qué tiene que ver con la calidad del diseño de muestreo que utilicemos el cálculo o estimación de la varianza?

Un indicador de la calidad del diseño se obtiene a partir del coeficiente de estimación de la varianza.

35.- ¿Cómo se relaciona en general un total, una media y una proporción?

Porque cuando la variable de interés y es dicotómica (i.e. sólo puede tomar valores 0 y 1), la expresión para computar la media poblacional:

$$\bar{y_u} = \frac{total}{tamaodelapoblacin} = \frac{t}{N} = \frac{\sum_u y_k}{N},$$

funciona también para calcular una proporción (i.e. la proporción de '1' en la población). Bajo este esquema, dado que una media y una proporción son virtualmente idénticas, podemos replantear sus estimaciones en función al total que se divide entre la constante N.

36.- Si la calidad de un estimador, una de las características de las que depende es el sesgo de éste, ¿Qué significa que un estimador sea insesgado formalmente hablando? ¿Y que significa en palabras coloquiales como las entendería para un político o cliente comercial?

La esperanza del estimador es el parámetro que está estimando. La explicación a un político sería: Imagínese que tiene un hijo, habrá algunas cosas en las que se parece a usted y otras en las que se parezca a su mamá, de alguna manera su hijo tiene ciertas características de usted que lo representa, un estimador insesgado es como si se hubiera clonado, es decir, ese hijo clon lo representa totalmente a usted.

37.- ¿Es lo mismo hablar del sesgo de un estimador que de que una muestra tiene sesgo, como habla coloquialmente la gente ajena a técnicas de muestreo? ¿Sí, no? Explique ampliamente.

No, hablar del sesgo como propiedad del estimador implica hablar de qué tanto difiere su valor esperado (o esperanza) y el valor verdadero del parámetro que intenta estimar; en cambio, cuando se emplea coloquialmente el término "sesgo" para hablar de si una muestra puede, o no, estar sesgada, se hace referencia al grado con que esa muestra, de acuerdo a su recolección, captura de manera "justa" la variabilidad contenida en la población (por ejemplo, levantar una encuesta en la Condesa y pretender extrapolar conclusiones para describir a toda la Ciudad de México).

38.- ¿Por qué formalmente hablando no existe una estimación insesgada?

Las estimaciones (a diferencia de los estimadores) representan los valores puntuales computados a partir del análisis de una muestra determinada, y de acuerdo con el Teorema del Límite Central, las estimaciones que se pueden obtener a partir de las distintas muestras posibles a extraer de la población varían de acuerdo a una distribución Normal. Por definición, el sesgo se evalúa según la discrepancia entre la media de mi estimador (la media de las posibles estimaciones a obtener) y el valor poblacional real que se intenta evaluar. En otras palabras, no tiene sentido hablar de estimaciones sesgadas en tanto que estas representan una sola observación de la distribución del estimador; por el contrario, sólo tiene sentido hablar del sesgo en términos de esta distribución as a whole.

39.- ¿Explique cómo se construye una distribución muestral de un estimador? Explique como para un chavito de preparatoria.

Supongamos que te interesa estimar la estatura promedio de los hombres que estudian en tu prepa; De acuerdo a lo que has visto en tu clase de Estadística, una buena forma de aproximarte a responder esta pregunta sería tomar "una muestra" (seleccionar a algunos de los hombres que estudian en tu escuela), medir sus estaturas y calcular el promedio. Digamos que bajo esta lógica, seleccionas al azar 50 estudiantes de la lista de alumnos inscritos a tu escuela, luego mides su estatura, calculas su promedio (X), y respondes así tu pregunta inicial: X es tu estimación sobre el promedio de la estatura de todos los estudiantes. Pero…¿y si en lugar de seleccionar a esos 50 estudiantes, hubieras elegido un grupo distinto?¿Habrías llegado al mismo valor de X? La respuesta es que muy probablemente no. Resulta que X, el "estimador" elegido para aproximarnos al promedio de las estaturas de nuestra población objetivo (el conjunto total de estudiantes hombres en tu escuela), nos llevará al cómputo de valores puntuales distintos, dependiendo de quiénes sean los estudiantes que incluyamos en la muestra. Decimos entonces que nuestro estimador X es una variable aleatoria, que toma distintos valores dependiendo de cuál sea la muestra con la que estamos trabajando en un momento dado. ¡Pero espera, hay más! Resulta que si tuvieras el tiempo y la pericia de realizar una cantidad enorme de muestras (digamos, repetir el proceso de seleccionar, medir y promediar unas 100 veces), y computaras un promedio por cada una de ellas, e hicieras un histograma para representar la frecuencia con que observas distintos valores de X, terminarías encontrando algo muy parecido a una distribución Normal, cuya media debería coincidir con el valor verdadero del promedio de las estaturas en toda la escuela.

40.- Hasta lo que hemos visto, si se quisieran mejorar las estimaciones. ¿En qué elementos tengo control (es decir, no depende del azar) y qué cosa usted podría alterar o mejorar?

Contar con el mejor marco muestral posible. Realizar un muestreo estratificado, cuidando elegir una variable de estratificación confiable. Incrementar, en la medida de lo posible, el tamaño de la muestra.

41.- ¿En poblaciones finitas, es posible determinar todas las muestras posibles? ¿Sirve de algo eso en la práctica, necesito listarlas todas?

Dependiendo el tamaño de la población, es posible que yo pueda definir el conjunto de todas las muestras posibles (por ejemplo, si mi población objetivo es el grupo de estudiantes

Sección 1 continúa en la siguiente página…

al cual le doy clases y sobre los cuales me interesa conocer el nivel de logro en la adquisición de cierto aprendizaje, a partir de las puntuaciones obtenidas en un examen). No obstante, si mi población finita es muy grande (por ejemplo, si hablamos de todas aquellas personas que desempeñan funciones docentes en el país), es poco probable que pueda identificar todos los elementos que componen mi muestra (N), y por tanto, el conjunto de posibles muestras a extraer. Asumiendo que N fuera conocida, el número de posibles conjuntos (muestras) a formar a partir de mi población (tomando en cuenta tanto el conjunto vacío como el censo) está dado por la expresión 2^N.

42.- ¿Para qué nos sirve el coeficiente de variación estimado? Explique su utilidad práctica a un subalterno que estudió matemáticas.

El coeficiente de variación estimada (CVE) funciona como una medida para evaluar el error estándar observado a la luz del estimador. Es decir, más que concentrarnos en el valor puntual del error, interesa que éste sea evaluado en función a aquello que está midiendo ($\hat{\theta}$). En este sentido, el CVE nos puede ayudar a distinguir cuál de varios posibles estimadores tiene una mayor calidad en sus estimaciones, en función de la razón entre su varianza y su media.

43.- ¿Cómo explicarle a un político o a un niño en términos coloquiales en realidad qué hace el coeficiente de variación? Ejemplifique si lo considera pertinente.

El coeficiente de variación te dice que tan variada es tú población, lo que buscamos es que tienda a 0 porque de esa manera la variación es más pequeña. Es como cuando juegas a los quemados en la escuela, es más fácil que le pegues a alguno cuando están todos amontonados (Cuando el coeficiente es cercano a 0) que cuando están todos dispersos en el patio. Así también depende si son muchos o pocos, si son muchos no importa a donde avientes la pelota, pero si son pocos, te conviene que todos estén en un solo lugar. Es decir el coeficiente de población tiende a 0 cuando tienes a muchos o cuando a los que tienes los tienes en un espacio muy pequeño.

44.- En palabras, sin fórmulas ni notación matemática. . . ¿De qué se trata el uso de los estimadores π o de Narain-Horvitz-Thompson? ¿Cuál es la idea intuitiva que hay detrás? Explique ampliamente demanera simple. Ejemplifique si lo considera pertinente.

Los estimadores de π de Narain-Horvitz-Thompson también son conocidos como estimadores de expansión simple, y parten de la idea de que los valores obtenidos en la muestra (y_k) deben ser "π-expandidos", es decir, relativizados en función a su propia probabilidad de inclusión en muestra. Por ejemplo:

$$\check{y_k} = \frac{y_k}{\pi_k}$$

45.- ¿Qué restricciones hay en las probabilidades de inclusión para poder utilizar los estimadores de Narain-Horvitz-Thompson? ¿Qué restricciones tengo para establecerlas?

La restricción es que toda la población tenga una probabilidad positiva de ser seleccionados para una muestra. Al momento de estimar la varianza, obliga a que cada pareja debe de tener una probabilidad positiva porque las probabilidades de inclusión (π_k) suelen ser usadas como denominador de las expresiones a trabajar.

46.- ¿Qué es la fracción de muestreo y qué información me da si la tengo términos porcentuales?

Es la proporción de unidades elegidas para la muestra, en porcentaje, nos estaría indicando el tamaño de la muestra con relación al tamaño de la población.

47.- Explique ¿qué significa estratificar en términos prácticos y en términos matemáticos?

En términos prácticos nos permite planear la logística de levantamiento de información dependiendo de la estratificación que hagamos, por contexto, geográfico etc. También nos permite, en caso de que la estratificación hubiera salido mal, quitar los estratos. Matemáticamente provoca particiones de la población que se pueden trabajar de manera individual.

48.- ¿Por qué se recomienda estratificar como una técnica útil para mejorar estimaciones? ¿Cómo convencería a su jefe ignorante en muestreo sin tanto tecnicismo?

Estratificar no compromete la precisión de las estimaciones que se puede. Es decir, incluso en el peor de los escenarios, asumiendo que la estratificación no se haya realizado de forma adecuada, las estimaciones obtenidas serán muy cercanas a aquellas que habría obtenido sin hacer la estratificación.

49.- Suponga que tiene un marco muestral de 40mil registros. Usted sabe de antemano que la variable Z, disponible en su marco, es "ideal" para utilizarse como variable de estratificación. esafortunadamente, no todos los registros en su marco tienen registros de esa variable. Aproximadamente un 20$\%$ de su marco muestral no presenta información sobre tal variable. ¿Cuál es la mejor alternativa que usted sugeriría? Discuta ampliamente las otras alternativas y por qué lo que propone es mejor. Convenza al jefe que estudió medicina.

Estimado jefe: de acuerdo con el trabajo que quiere realizar le tengo dos propuestas.

1. Las personas sin la variable en el marco muestral, definirlos como un estrato.
2. No estratificar y no utilizar esa variable.

La opción dos implicaría que la “variable” que podría darnos amplios conocimientos no se tome en cuenta y se utilice todos los registros. Sin embargo la segunda opción nos permite utilizar todos los registros y además incluir esta “variable”, lo cual nos permitiría tener estimaciones para toda la población.

Sección 2 VERDADERO O FALSO (con justificación)

50.- La función diseño de muestreo es la que determina las propiedades estadísticas del estadístico que estoy utilizando como estimador.

Sección 2 continúa en la siguiente página...

Verdadero. Una vez definida $p(s)$, cualquier muestra observada constituye una realización de la variable aleatoria (s). La función diseño determina las propiedades estadísticas de los estimadores que se quieran computar a partir del ejercicio de muestreo.

51.- En muestreo directo de elementos, es decir en 1 etapa, y bajo un diseño SI se requiere forzosamente tener el marco muestral completo que identifique a los elementos de la población.

Verdadero: Debido a que es necesario tener un marco muestral para realizar la selección probabilística.

52.- Si se incorporan más etapas al diseño de muestreo regularmente se aumenta la varianza del estimador.

Verdadero: Si partimos de la idea que para mejorar un diseño mejoran los estimadores, el incorporar más etapas lo empeoraría, en el caso de la varianza, aumenta las fuentes de variabilidad.

53.- La ventaja principal de las muestras probabilísticas sobre las no probabilísticas es que no hay errores no muestrales.

Falso. Los errores no muestrales son ajenos al diseño muestral (por ejemplo, las características del instrumento que se esté utilizando para extraer información de la muestra)

54.- Para mejorar la precisión en un diseño de muestreo de varias etapas se sugiere tratar de aumentar el tamaño de muestra de las unidades primarias de muestreo, es decir el número de elementos a muestrear en la primera etapa. Muchas veces esto tiene que hacerse disminuyendo el número de unidades últimas de muestreo para no afectar el tamaño de muestra global.

Verdadero, siempre y cuando el tamaño de muestra sea fijo.

55.- Es posible obtener muestras insesgadas incluso bajo diseños de muestreo diferentes al SI.

Falso. El insesgamiento es una característica de los estimadores no de las muestras.

56.- El tamaño de muestra se determina mayormente por el tamaño de la población objetivo.

Falso. El principal factor para determinar el tamaño de muestra es la parte económica y el nivel de precisión.

57.- En un muestreo SI. Si censamos se obtiene una varianza del estimador igual a cero y también la estimación de la varianza del estimador es igual a cero.

Verdadero. En el censo el estimador es igual al parámetro

58.- Una proporción es una media de variables continuas.

Falso. Sólo funciona con variables dicotómicas.

59.- En el muestreo aleatorio simple, todas las muestras tienen la misma probabilidad de ser extraídas.

Verdadero. La extracción de muestras mediante un mecanismo aleatorio que permita que haya equiprobabilidad en la extracción de cada muestra s posible, constituye la cuarta condición para poder decir que se está trabajando con muestreo probabilístico.

60.- En el muestreo aleatorio simple estratificado, todos los elementos de la población tienen la misma probabilidad de ser seleccionados.

Falso. Esto debido a que cada estrato tiene su propia probabilidad de inclusión.

61.- En el muestreo aleatorio simple, todos los elementos de la población tienen la misma probabilidad de ser seleccionados.

Verdadero (al menos en teoría, porque el que se cumpla o no este supuesto depende del proceso que se decida implementar como parte del esquema de selección). El propósito del Muestreo Aleatorio Simple es la extracción, sin reemplazos, de n elementos dentro de una población de tamaño N, procurando que cada selección se haga con la misma probabilidad. Por ejemplo, se selecciona al primer elemento con una probabilidad de $\frac{1}{N}$, al segundo elemento con $\frac{1}{N-1}$ y al k-ésimo elemento, con una probabilidad de $\frac{1}{(N-k+1}$; o bien, puedo seleccionar todos los elementos con la misma probabilidad y "manualmente" descartar aquellos que aparezcan repetidos, y seguir con el proceso de extracción hasta acabar con una muestra de tamaño n.

62.- Para mejorar la precisión en un diseño de muestreo se sugiere aumentar el tamaño de muestra.

Verdadero. Debido a que disminuye la variabilidad, como se le explicó al chico de la pregunta 43. Aunque hay otras alternativas para mejorar la precisión.

63.- Siempre que tenga un nivel de precisión en los dominios de estimación, al combinar las estimaciones para dar una estimación global, el nivel de precisión de la estimación global es mejor que el de la estimación por dominios.

Verdadero. Ya que las estimaciones son particiones de n, los elementos son menores, de tal manera que no siempre cumplen con la precisión que se desearía.

64.- Para estimar proporciones se pueden usar prácticamente las mismas expresiones matemáticas que para estimar medias.

Verdadero para variables dicotómicas.

65.- El muestreo polietápico, es decir en más de dos etapas de muestreo requiere forzosamente de un marco muestral completo que identifique a todas las unidades últimas de muestreo.

Falso. Una de las razones de utilizar etapas es para remediar problemas con el marco muestral.

66.- El deff teórico para cualquier estimador del diseño SI es igual a cero siempre. Esto por su definición.

Falso. El deff en SI es igual a 1

67.- Siempre que utilizamos conglomeración se aumenta la precisión en mis estimaciones.

Falso. De hecho, ocurre todo lo contrario: conforme se agregan nuevas etapas al muestreo, se pierde cada vez más precisión. En general, esta práctica sólo es recomendada cuando no se cuenta con un marco muestral apropiado, o bien, cuando se quiere "economizar" el diseño muestral.

68.- Según la teoría vista en el curso. El esquema real de muestreo puede ser diferente a mi función diseño de muestreo al momento de estimar. Se vale y es correcto.

Falso. La función de diseño representa la base formal o matemática que, al definir una distribución de probabilidad para describir la variable aleatoria S (el conjunto de todas las muestras posibles), determina las propiedades estadísticas de las cantidades aleatorias a calcular. El esquema real de muestreo refiere al protocolo "real" que guía el proceso de selección de la muestra (si se va a usar un diseño estratificado, con conglomerados, con las mismas probabilidades de inclusión para todos los elementos o no, etc.). Para dar validez al proceso de inferencia, es necesario que ambos elementos mantengan una relación estrecha; es decir, que el esquema de muestreo sea congruente con la función diseño propuesta.

69.- Siempre que se quiera mejorar la precisión en un diseño demuestreo en varias etapas se sugiere reducir el número de etapas, es decir dejar de conglomerar para algunas etapas.

Verdadero. La mayor parte de la varianza se captura en el primer desgloce (etapa I), de acuerdo con el principio del conglomerado último

70.- El coeficiente de variación (teórico, no estimado) puede tener valores iguales a cero si censo.

Verdadero (Por ejemplo, si la población tiene tamaño 1)

71.- El error estándar y la desviación estándar no son lo mismo.

Verdadero: la desviación estándar es una medida general de dispersión en cualquier variable aleatoria, pero el error estándar específicamente se refiere a la desviación estándar

Sección 2 continúa en la siguiente página...

de la distribución muestral de cierto estimador.

72.- Si muestreamos bajo el enfoque basado en modelos lo estocástico o variabilidad está en el componente aleatorio del modelo.

Verdadero. El modelo que se elige para describir la "superpoblación" generadora de nuestra población objetivo, es quien señala qué componentes son aleatorios y cuáles son determinísticos.

73.- No se pueden calcular errores de estimación con muestreo no probabilístico. Por eso no tiene sentido calcular un tamaño de muestra.

Verdadero. Bajo el enfoque basado en diseño, no tiene sentido calcular los errores o el tamaños muestra.

74.- Para calcular un tamaño de muestra a cierta precisión y confianza necesito siempre el supuesto de Normalidad.

Falso. No es necesario el supuesto de normalidad, ya que por medio de otros métodos se puede hacer el cálculo, por ejemplo, utilizando el coeficiente de variación.

75.- Una manera de estimar a N, el tamaño de la población, es sumando los factores de expansión de los individuos caídos en muestra

Verdadero. Por definición los factores de expansión están en función del número de individuos que representan en una población.

76.- La probabilidad de inclusión conjunta para el par de elementos (k, k), es igual a la probabilidad de inclusión de primer orden del elemento k.

Verdadero, porque $\pi_{kk} = Pr(I_k^2 = 1) = Pr(I_k = 1) = \pi_k$

77.- Es posible tener probabilidades de inclusión de primer orden igual a $\frac{n}{N}$ y tener un diseño de muestreo $p(\cdot)$ distinto del muestreo SI.

Falso. El muestreo SI es precisamente aquel enfoque que busca que todos los elementos de la población tengan la misma probabilidad de inlcusión, lo cual se cumpliría si todos los elementos tienen π_k igual a $\frac{n}{N}$, ya que busca que todos los elementos tengan la misma probabilidad de inclusión.

78.- No se puede estimar puntualmente una proporción si no se conocen sus probabilidades π_{kl}.

Falso. Se puede realizar la estimación puntual a partir de las probabilidades de inclusión de primer orden.

79.- Con las expresiones que vimos en clase, no es posible calcular la varianza con un tamaño de muestra menor a 2.

Falso. La varianza de una muestra menor a 2 sería simplemente 0. Es por ello que, cuando estamos hablando del tamaño de muestra en un estrato, se recomienda "colapsar" el estrato con $n_h = 1$ con el que más se le parezca, en términos de la variable de estratificación.

80.- Las probabilidades de inclusión de primer orden son iguales a las probabilidades de inclusión conjuntas si trabajamos con un diseño de muestreo aleatorio simple.

Falso. Para obtener ambas razones se emplean fórmulas distintas, n/N para primer orden y $\frac{n(n-1)}{N(N-1)}$ para segundo orden.

81.- El tamaño de muestra se determina mayormente por el tamaño de la población objetivo.

Es la misma pregunta que la 56, aunque sigue siendo falso.

82.- Es conservador que la estimación de varianza de un estimador tenga un sesgo negativo a uno positivo. Es decir, es conservador obtener errores estándares ligeramente sub-estimados.

Falso. El subestimar los errores estándar nos estaría dando una estimación errónea.

83.- En las expresiones de estimación puntual de Narain-Horvitz-Thompson las probabilidades de inclusión pueden ser arbitrarias sin restricción.

Falso. Como se dijo en la pregunta 45 una de las limitantes es que tienen que ser positivas

84.- Los errores no muestrales siempre son pequeños en comparación a los errores muestrales.

Falso: si tenemos un instrumento de medición defectuoso que devuelve medidas extremadamente variables en situaciones controladas, por ejemplo, podemos esperar que al aplicarlo en cierta muestra dicho error de medición tenga más peso que el error muestral.

85.- Al incorporar más etapas al diseño de muestreo se puede perder el insesgamiento del estimador puntual lineal.

Falso. El número o nivel de etapas no afecta la calidad del estimador.

86.- Siempre que la población es mucho más grande, la muestra tiene que ser mucho más grande.

Falso. El tamaño de la muestra tiene que ver con elementos de precisión o económicos, más que con el tamaño de la población.

87.- Siempre que se quiera mejorar la precisión en una etapa específica de muestreo se sugiere disminuir el número de unidades muestrales correspondientes a esa etapa.

Falso. Entre el 90 y el 95% del error de medición muestral se encuentra contenido en la primera etapa, por lo que, si queremos mejorar la precisión en la primera etapa debemos incrementar el número de unidades muestrales primarias. Por otro lado, si estamos hablando de una etapa posterior (la segunda, o la tercera), entonces sí, lo recomendable es extraer un número menor de unidades muestrales secundarias o terciarias, a cambio de muestrear más en la primera etapa.

88.- Se necesitan al menos tanta cantidad de estratos como cantidad de dominios de estudio tengo planeados.

Verdadero: Los dominios constituyen subconjuntos de la población acerca de los cuales interesa poder decir algo, (o bien, a lo largo de los cuales interesa distinguir/diferencial el ejercicio de inferencia) y suelen ser planificados. Por su parte, los estratos constituyen una estrategia de segmentación de la población en subgrupos definidos a partir de una variable de estratificación que resulte relevante para nuestra variable de interés, (por ejemplo, si lo que interesa es evaluar la variabilidad de las habilidades matemáticas en mi población, yo podría hacer una muestra estratificada a partir de la cual yo recolecte elementos con distintos niveles de logro en algún exámen de matemáticas que yo pueda usar como mi estándar de oro). En este sentido, sí, tengo que tener "al menos" tantos estratos como dominios (de forma que yo pueda generar inferencias diferenciadas por cada dominio), en el entendido de que podría tener más estratos que dominios, (de manera que dentro de los dominios, yo esté trabajando con estratos).

89.- Si censamos una población de elementos tenemos una fracción de muestreo de 1.

Verdadero: La fracción de muestreo se define como $\frac{n}{N}$. Al hablar de un censo, queda implícito que $n = N$, así que sí, la fracción de muestreo es 1.

90.- De acuerdo a la teoría vista en el curso. El total de elementos en mi población a los que les asigno probabilidad $\pi_k = 1$ no puede ser mayor al tamaño de muestra n.

Verdadero. Al ser 1 esto indica que el individuo será incluido en la muestra, por lo que no puede existir más individuos que la muestra.

91.- Si sumamos las probabilidades de inclusión de los elementos en toda mi población obtenemos exactamente el valor n.

Falso. Esto aplica sólo cuando el tamaño de muestra es fijo, sin embargo, se puede obtener el tamaños esperado de la muestra.

92.- Cuando usamos muestreo aleatorio simple no podemos asumir el gran supuesto estadístico de tener observaciones independientes idénticamente distribuidas.

Depende de cómo se implemente: en general, teóricamente la afirmación es falsa (i.e., sí se pueden asumir vaiid), pero en la práctica la afirmación es verdadera (i.e., la forma de muestrear individuos no permite asumir que son iid). Para que el supuesto de observaciones iid pueda defenderse, sería necesario muestrear personas con reemplazo, es decir, tomar un individuo al azar de acuerdo con su probabilidad de inclusión, medirlo, y "regresarlo" a la población, de tal forma que la misma persona pueda ser incluida más de una vez en la muestra. En la práctica, sin embargo, este procedimiento rara vez se sigue y lo más común es incluir únicamente una vez a cada persona en la muestra. Como en la práctica las personas son muestreadas sin reemplazo, la probabilidad de inclusión de cierto individuo no es la misma en todas las etapas de recolección de la muestra, violando los supuestos de independencia y de distribución idéntica.

93.- Siempre que la población es más chica mejora la precisión de mis cálculos.

Falso. El tamaño de la población no es un factor que afecte la precisión de la medición.

94.- Por su definición, Δ_{kl} es la correlación de las indicadoras de inclusión muestral de los elementos k y l.

En realidad Δ_{kl} refiere a la covarianza de las indicadoras de inclusión muestral de los elementos k y l.

95.- Un parámetro tiene variabilidad y esta se mide por la varianza de éste, pero para calcular su varianza se requiere de toda la información de la población.

Falso. Los parámetros no tienen variabilidad, son valores fijos que representan alguna característica de la población; los estimadores que se eligen para estimar dichos valores, sí.

96.- Cuando alcanzo cierto error estándar en mis estimaciones globales, si quiero dar resultados por sub-poblaciones, dominios o cruces, estos tendrán un error estándar más grande.

Falso. Como cada subpoblación es individual, el error estándar en algunas de ellas puede ser menor a la global.

97.- Las probabilidades de inclusión de primer orden son iguales a las probabilidades de inclusión conjuntas si trabajamos con un diseño de muestreo aleatorio simple.

Al hablar de Muestreo Aleatorio Simple se hace referencia a un esquema de selección de la muestra donde se extraen n (conocida) elementos de una población de tamaño N sin reemplazo, siendo que cada selección sea equiprobable. Por tanto, se tienen una probabilidad de inclusión de $\frac{n}{N}$ para cualquier elemento k contenido en la población, y una probabilidad de inclusión conjunta $\frac{n(n-1)}{N(N-1)}$ para dos elementos distintos.

98.- Las probabilidades de inclusión de primer orden son iguales a las probabilidades de inclusión conjuntas si trabajamos con un diseño de muestreo Bernoulli.

Falso. En el muestreo Bernoulli se asigna una misma probabilidad de inclusión a todos los elementos contenidos en la población ($\pi_k = \pi$), y dado que los eventos $k \in s$ y $l \in s$ son independientes, la probabilidad de inclusión conjunta (o de segundo orden $\pi_k l$) estaría determinada por π^2

99.- Siempre que se quiera mejorar la precisión en un diseño de muestreo se sugiere estratificar.

Verdadero. Si los estratos son definidos en función a una variable de estratificación relevante o intrínsecamente relacionada a la variable de interés, el muestreo estratificado resultante estaría recuperando en mayor medida la variabilidad de la variable de interés en la población.

100.- La varianza del estimador de un parámetro en un muestreo estratificado aleatorio simple es casi siempre menor que la varianza si no hay estratos y se utilizó un muestreo aleatorio simple.

Verdadero: en general la varianza bajo muestreo estratificado es menor que en muestreo no estratificado porque las unidades dentro de los estratos se parecen entre sí (p.ej., si un estrato es "población que vive en playa" y otro "población que vive en montaña", podemos esperar que las personas dentro de cada grupo de población sean muy parecidas en la característica: "resistencia física/cardiovascular", aunque los dos grupos difieran marcadamente en ella). Como las unidades que conforman los estratos son homogéneas, la precisión del estimador dentro de cada estrato es alta, es decir, podemos tener mucha certeza sobre cuál es el valor de la característica dentro de cada estrato, y en tanto que la estimación de la característica en la población completa depende de la certeza a nivel estrato, también podemos estar muy seguros sobre cuál es el valor de la característica en la población. En el peor de los casos, si los estratos no reflejan grupos homogéneos, la precisión de la estimación en la población corresponde con la precisión que se hubiera obtenido de no haber estratificado.

101.- Siempre se disminuye la varianza del estimador si se aumenta el tamaño de muestra en un diseño SI.

Verdadero. Al aumento de la muestra es inversamente proporcional a la varianza del estimador, esto dado por el teorema central del límite.

102.- De acuerdo al curso. No es posible asignar probabilidades de inclusión 1 a algunos elementos en el marco muestral porque no estaríamos haciendo muestreo probabilístico.

Falso. Es factible que las probabilidades de inclusión sean arbitrarías, aunque podrían tener un impacto negativo en el diseño de muestra.

103.- Si estratificamos un diseño de muestreo (sin importar si es un diseño de muestreo de más de una etapa), ésta puede hacer perder al estimador lineal su insesgamiento.

Falso. Como se colocó en la pregunta 55 el insesgamiento es una característica de los estimadores, no de la muestras, por lo tanto, esta no depende de la estatificación o número de etapas.

104.- No es posible tener tamaño de muestra 1 en un estrato, aun cuando su tamaño poblacional sea 1.

Falso. El único caso donde yo puedo observar $n_h = 1$ es cuando uno de los estratos tiene un tamaño poblacional de 1, en cuyo caso estaría haciendo una extracción censal del mismo.

105.- No existen restricciones en el tamaño de muestra asignado a los estratos cuando se incorpora una estratificación al diseño de muestreo utilizado.

Falso. n_h no puede ser menor a 2 (a menos que el tamaño poblacional de ese estrato sea 1 y esté censando el estrato).

106.- En un muestreo en varias etapas. No es posible utilizar la muestra de la etapa anterior como población para extraer muestras en la etapa siguiente.

Falso. Sí es posible utilizar la muestra de la etapa anterior como población de referencia. De hecho, las probabilidades de inclusión π_k se determinan por cada etapa

Sección 3 DESARROLLO ALGEBRÁICO

Vimos en clase (y usted demostró como tarea opcional) que:

$$E(I_k) = \pi_k$$

$$V(I_k) = \pi_k(1 - \pi_k)$$

$$C(I_k, I_l) = \pi_{kl} - \pi_k\pi_l =^{def} \Delta_{kl}$$

Sea n_s el tamaño de muestra para cualquier diseño de muestreo, tenemos que éste puede expresarse en términos de las indicadoraas de inclusión muestral I_k como: $n_s = \sum_u I_k$

(a) Calcule $E(n_s)$

Partiendo de la linealidad de la esperanza, se obtiene que:

$$E(n_s) = E(\textstyle\sum_u I_k) = \sum_u E(I_k) = \sum_u \pi_k$$

(b) Complete la expresión para $V(n_s)$*, rellenando los espacios, sabiendo que:*

$$V(\textstyle\sum_U I_k) = \sum_{k\in U} \sum_{l\in U} C(I_k, I_l)$$

Usamos la identidad sugerida para dividir la doble suma en dos casos: Primero cuando $k = l$ y luego, cuando son distintos. De esta forma se tiene que:

$$V(n_s) = V(\sum_U I_k) = \sum_{k\in U} \sum_{l\in U} C(I_k, I_l)$$
$$= \sum_{k\in U} C(I_k, I_k) + \sum_{k\in U} \sum l \in UC(I_k, I_l)$$

El motivo detrás de la separación d estos casos es que podemos identificar la suma que aparece en la expresión a completar como la suma de las varianzas de las indicadoras I_k. Es decir, notamos que $C(I_k, I_k) = V(I_k) = \pi_k(1 - \pi_k)$. Además, para k distinto de l, se tiene que $C(I_k, I_l) = \Delta_{kl}$, por lo que al sustituir en la expresión anterior se obtiene:

$$V(n_s) = \sum_U \pi_k \cdot (1 - \pi_k) + \sum_{k\in U} \sum_{l\in U} \Delta_{kl}$$

Lo que completa la primera expresión. Para la segunda, sustituimos $C(I_k, I_l)$ dentro de la doble suma para obtener:

$$V(n_s) = \sum_{k\in U} \sum_{l\in U} (\pi_{kl} - \pi_k \cdot \pi_l) = \sum_{k\in U} \sum_{l\in U} (\pi_{kl}\pi_{kl} - \sum_{k\in U} \sum_{l\in U} \pi_k \cdot \pi_l$$

Ahora bien, la doble suma que estamos restando es el producto de los pares de valores π_k; esto lo podemos factorizar como $\sum_{k\in U} \sum_{l\in U} \pi_k \cdot \pi_l = (\sum_{k\in U} \pi_k)^2$.

Por otro lado, dado que las probabilidades de inclusión de segundo orden coinciden con las de primer orden para $k = l$, podemos separar la doble suma de π_{kl}, en estos dos casos nuevamente para obtener:

$$V(n_s) = \sum_{k\in U} \pi_k + \sum_{k\in U} \sum_{l\in U} \pi_{kl} - (\sum_{k\in U} \pi_k)^2$$