Examen Parcial - Tópicos

Grupo 5

Claudia Mirela Vivas Alejandro
Sonia Rosmery Asto Mercado
Maria Erika Yangali Araujo
Andrea Pamela Salazar Zapata

October 2022

1. Parte 1 - Conceptos clave y temas de frontera

1. Responda dos de las siguientes preguntas. La respuesta a cada pregunta no debe pasar de 75 palabras.

   a. *¿En qué se diferencian un dominio y un estrato?*

   Tanto el dominio como el estrato son divisiones de la población objetivo de estudio. Sin embargo, se diferencian debido a los objetivos de su uso. Los dominios indican el nivel de representatividad y por tanto inferecia que tendrán nuestros estimadores (nivel de regiones, por ejemplo), los dominos se pueden definir a nivel geográfica (región, distrital, etc) o subpoblaciones (comunidad LGBT, adultos mayores, etc). Por otra parte, el estrato es información adicional que proviene del marco muestral que permite reducir la varianza de los estimadores.

   b. *Explique en sus propias palabras que es el efecto de diseño (DEFF)*

El efecto del diseño es una razón entre la varianza del muestreo diseñado (estratificado, por conglomerados o muestreo complejo) con la varianza del muestreo aleatorio simple (MAS) para un mismo tamaño de muestra. De esta manera, podemos saber que cuánto aumenta o reduce la varianza debido al diseño del muestreo. Por ejemplo, para un MA, el deff es menor o igual a 1, pero para un muestreo aleatorio por clusters es mayor a 1.

2. Elija un video del PAIDEIA en donde ningún miembro de su grupo haya participado. Elabore un comentario (hasta 400 palabras) que incluya:

   a. Comentario general sobre el tema abordado en el video

   b. Una crítica constructiva a cómo se abordó el tema o elementos que Ud. considera que faltaron.

   c. Explique qué se aborda en algún documento del material citado en las láminas de extensiones/otros temas y por qué resulta (o no) interesante. Si no hay esta lámina para el tema que eligió ¿cómo complementaría la exposición? Use referencias académicas.

   d. Importancia y un ejemplo hipotético de cómo se podría aplicar este tema a un problema de política en el Perú.

      El método de Small Area Estimation (SAE, en adelante) es importante pues como mencionan en el video, existe una necesidad de obtener indicadores o información de subgrupos de poblaciones pequeñas y el SAE surge como una respuesta ante esta necesidad de desagregaciones. Si bien durante la presentación del tema se abordó la descripción, los métodos y dentro de estos se enfocaron en los estimadores, consideramos que se debió ejemplificar a profundidad a que se refiere con áreas pequeñas o a qué tipo de poblaciones se refiere. Asimismo, para una persona que no sabe nada del tema es un poco complicado entender la metodología del SAE, si solo nos enfocamos en el desarrollo matemático de los estimadores. A pesar de que mencionaron las desagregaciones que se necesita para los Objetivos de Desarrollo Sostenible, consideramos que se debió profundizar más en dicho ejemplo, el porqué es importante las

desagregaciones para la ODS.

Por otro lado, en el material académico entregado del SAE, se señala la existencia de modelos adicionales como los modelos de efectos aleatorios o modelos de correlaciones temporales entre variables. Sin embargo, no se da una mayor explicación sobre estos modelos (en qué situaciones utilizarlas, ventajas y desventajas), por lo que deja de resultar interesante la mención de dichos modelos en la presentación . Respecto a ello, consideramos que la presentación podría ser complementada con información sobre los métodos de SAE que utiliza el INEI en el Perú como la metodología de Elbers, Lanjouw y Lanjouw (2003)[1] para la creación del mapa de pobreza en el Perú (INEI, 2020)[2] .

Por último, el SAE puede aplicarse a diversas encuestas en el Perú para generar información esencial sobre el desarrollo de políticas públicas. Un ejemplo hipotético es la evolución de la anemia en menores de 3 años en el Perú a nivel distrital. Sin embargo, la ENDES tiene una representatividad hasta el nivel departamental, por lo que el SAE es una técnica muy útil para construir este indicador de prevalencia. De esta manera, conoceremos en qué distritos debemos replantear políticas, ya que, si solo hacemos un análisis a nivel de departamento, podríamos tener una conclusión errónea, es decir, pensar que la anemia ha disminuido en el departamento e ignorar un posible aumento de la anemia en uno de sus distritos.

2. Parte 2 – Obtener muestras

Use la base del Censo Nacional de Población Penitenciaria (CENPE) 2016 de la página del INEI (http://iinei.inei.gob.pe/microdatos/). Puede descargar solo el módulo “CARATULA” o también más módulos de dicho censo.

1. Extraiga una muestra estratificada y por conglomerados (una etapa) que contenga entre

[1] Elbers, C., Lanjouw, J. O., Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. Econometrica, 71(1), 355-364

[2] Instituto Nacional de Estadística e Informática (2020). Mapa de pobreza monetaria provincial y distrital 2018

el 1 % y 5 % de la población (toda la base). Especificar qué variables se utilizan para los estratos y conglomerados y discutir por qué son adecuadas para dicho propósito. Discutir también si hay otras variables que podrían haberse utilizado para los estratos o conglomerados.

El usó un umbral de 2000 reos dentro de un centro penitenciario como estrato para el diseño, pues se observa heterogeneidad entre estratos y homogeneidad intraestrato. Además, se usó a los centros penitenciarios como variable para crear los conglomerados. Respecto a esta última variable, inicialmente se pensó en usar las áreas geográficas como conglomerados. Sin embargo, no se tenía precisión sobre la cantidad de centro penitenciarios que podían caer dentro de cada conglomerado, a diferencia de encuestas como la ENAHO, donde los conglomerados son áreas geográficas que albergan en promedio a 400 hogares. Asimismo, un segundo factor que consideramos para definir el conglomerado fue el domino, pues en la siguiente sección de la pregunta nos pedían realizar estadísticos a nivel de personas. Así que, escoger a los departamentos, provincias o distritos como conglomerados no iba a permitir que los estadísticos fuera representativos a nivel de personas, ya que en este ejercicio solo se permitió generar un diseño estratificado y conglomerado por una etapa asi que no se iba poder muestrear a unidades más pequeñas. Otro punto que consideramos en la elección de conglomerados y estratos fue la unidad de muestreo primaria (PSU), la cual hace referencia a la unidad de muestreo que se usa en la primera etapa. La unidad de referencia que usamos son los centros penitenciarios. En ese sentido, el estrato que definimos hace referencia a una característica de los centros penitenciarios y no a los reclusos, mantener un correlato de las unidades de muestreo entre el conglomerado y estrato es importante de forma que se pueda hacer una correcta aleatorización de los conglomerados por estrato.

Asimismo, se podría emplearse las variables geográficas como variables para definir los conglomerados, pero se observó que no existen centros penitenciaros en todos los distritos y provincias.Por ello, un conglomerado por macroregiones hubiera necesitado de una base de apoyo para definir las áreas geográficas. Por otro lado, una variable muy interesante que

podria haber sido usada como variable para la estratificación es el género. Con respecto a ello, al realizar una revisión de los estadísticos descriptivos del censo, se observó que el 94 % de reclusos son varones, mientras que las mujeres solo representan el 6 %. Uno de los objetivos de las muestras estratificadas es lograr usar información como la variable de género para realizar una muestra más representativa y reducir la varianza. Sin embargo, esta variable no fue considerada en nuestro diseño muestra, debido a que no tenía la misma unidad de muestreo que los conglomerados.

2. El porcentaje de personas que realizaron un delito contra la seguridad pública es de 26.8 %, sin embargo no tiene intervalos de confianza. No obstante, para verificar la validez de nuestros resultados se calculó el porcentaje de personas que comentieron delitos contra la seguridad pública en el censo, esta es de 25.35 % esto indica que el estadístico que obtuvimos no está lejos del estadístico poblacional .Asimimo, 1646 individuos cometieron delitos contra la seguridad pública.

3. Parte 3 – Uso de encuestas reales

A cada grupo se le asignará una encuesta que pueda ser descargada de la página de microdatos del INEI. Dicha encuesta será utilizada para este ejercicio.

1. Explicar el diseño muestral de la encuesta. Luego, declarar el diseño muestral (completo) en el software. Para ello, se recomienda revisar la ficha técnica de la encuesta.

   De acuerdo a la ficha técnica de la ENEVIC, el diseño muestral esta conformada por todas las personas mayores de 15 años que residen en áreas urbanas. Además, el marco muestral, es una cartografía e información estadística de los Censos Nacionales 2007: XI de población y VI de vivienda, con actualización del Empadronamiento de Población y Vivienda de 2012 y 2013. Por otro lado, el tipo de muestreo al ser probabilista se intuye que la muestra fue aleatoria. En ese sentido, es un muestreo de áreas, estratificada, multietápicas e independiente en cada ámbito de estudio. Además, cuando se menciona multiétapicas hace referencia a un muestreo realizado por etapas. En este caso se realizó el muestreo

por dos etapas, el cual tiene un nivel de confianza del 95 %. Asimismo, el tamaño de la muestra es de 39,840 viviendas particulares correspondiente de la capital de la región o departamento. A nivel de conglomerado de viviendas se trabaja con 4,980 conglomerados y dentro de cada conglomerado se trabaja con una muestra de 8 viviendas.

2. Construya dos variables nuevas de su interés: (i) una dicotómica, (ii) una continua

   Se creó una variable dicotómica de **seguro de salud** la cual asigna "1" si es que la persona tenía algún tipo de seguro específico y "0" si no posee seguro de salud. Para la variable continua, se generó la variable **número total de agresiones** donde sumamos la cantidad de agresiones que sufrió cada individuo. No obstante, los valores que toma esta variable no son muchos, lo cual podría llevar a pensar que es una variable categórica, sin embargo, esta variable es continua, pues los valores que adopta la variable no hacen referencia a un ordenamiento.

3. Obtenga la correlación intraclúster (ICC), el efecto del diseño (DEFF) y el tamaño de muestra efecto (effective sample size) para ambas variables. Explique ¿cómo se relacionan estos conceptos en sus cálculos? ¿cómo afecta los valores que ha obtenido a la precisión de sus estimados?

   En cuánto a la variable continua **número total de agresiones**, el ICC muestra un valor de 0.30977. Esto significa que las personas de un mismo clúster se parecen o poseen una similitud de un 30 %. Asimismo, al calcular el efecto del diseño para esta variable se muestra que la varianza del muestreo complejo (estratificado y por conglomerados) es 3.68 veces la varianza de un muestreo aleatorio simple. En linea con ello, a mayor efecto de diseño, el tamaño de muestra efectiva es menor. Por ello, al calcular el tamaño de muestra efecto, se encuentra que, si bien en la encuesta se ha entrevistado a 38 848, pero dado que las personas son similares entre los clusters es como si solo se hubiera entrevistado aproximadamente a 10 545 personas.

   Del mismo modo, respecto a la variable dicotómica **seguro de salud**, se encontró con un

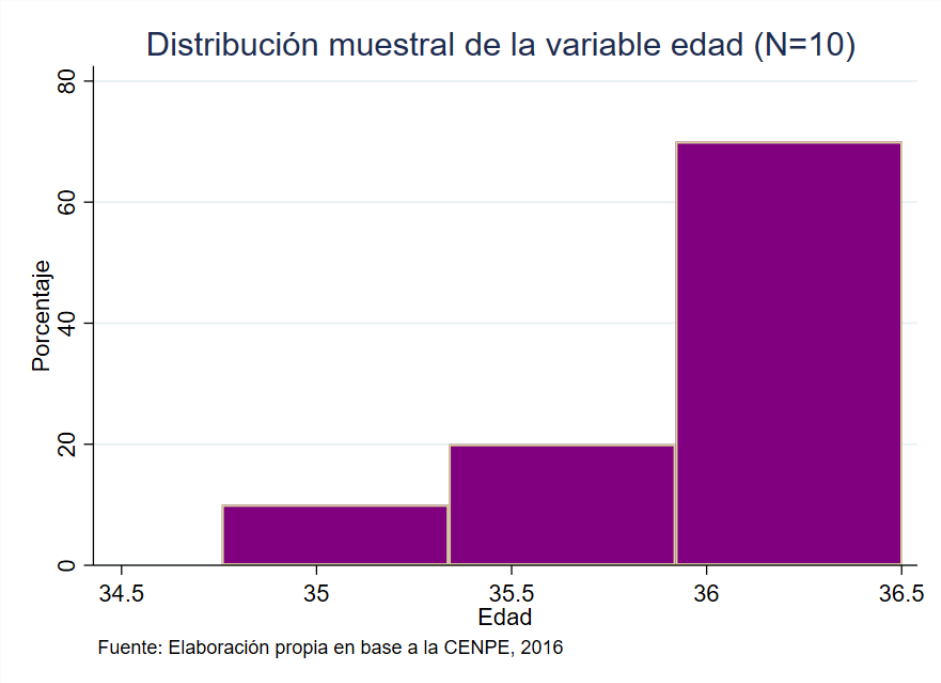
ICC de 0.07443. Esto significa que las personas de un mismo clúster poseen una similitud de un 7 %. En cuanto al efecto del diseño en la variable se encontró que la varianza del muestreo complejo (estratificado y por conglomerados) es 3.94 veces más que la varianza de un muestreo aleatorio simple. Finalmente, se realizó el tamaño de muestra efecto para esta variable dicotómica, encontrando que si bien se ha entrevistado a 38 848 es como si realmente se hubieran entrevistado a 9 859 personas mediante un muestreo aleatorio simple para obtener el mismo nivel de precisión. Es importante resaltar que tanto el ICC, efecto del diseño y el tamaño de muestra efectivo dependen de la variable de interes que deseamos analizar. Por ello, los valores obtenidos para las variables ”número total de agresiones” y ”seguro de salud” son distintas.

4. Parte 4 – Proof by Stata

”Demuestre” el Teorema Central del Límite y la Ley de los Grandes Números usando Stata. Esta demostración refiere a la distribución muestral de un Muestro Aleatorio Simple sin reemplazamiento. Para ello utilice la variable “EDAD” del módulo “CARATULA” del CENPE, una fracción muestral de 1 % y conduzca el ejercicio con 10, 100, 1000 y 10000 repeticiones. ¿Qué pasa cuando aumenta el número de iteraciones? ¿A qué valor converge el promedio muestral? ¿Qué distribución aproximada tiene la distribucion muestral? Para esta última pregunta presente el histograma de cada ejercicio, y corra por lo menos una prueba formal (i.e. Jarque-Bera, Shapiro-Wilks o Skewness-Kurtosis).

El teorema del límite Central nos menciona que para una muestra lo suficientemente grande, el promedio de una variable aleatoria, cuando tiende al infinito, la distribución de ese promedio es una normal. Por ello, se observa que a medida que crece el número de iteraciones la distribución de las medias se va acercando a una distribución normal. Asimismo, la ley de los grandes números nos menciona que el promedio de una muestra converge al valor esperado. En ese sentido, si se tiene una muestra lo suficientemente grande captaremos el verdadero valor del parámetro de esa población. Por ello, con el crecimiento de las iteraciones se observa que la media de las distribuciones se acercan a la media poblacional de la variable ”Edad” (36.03). De esta forma,

se demuestra que se cumple el teorema central del límite y la ley de los grandes números.

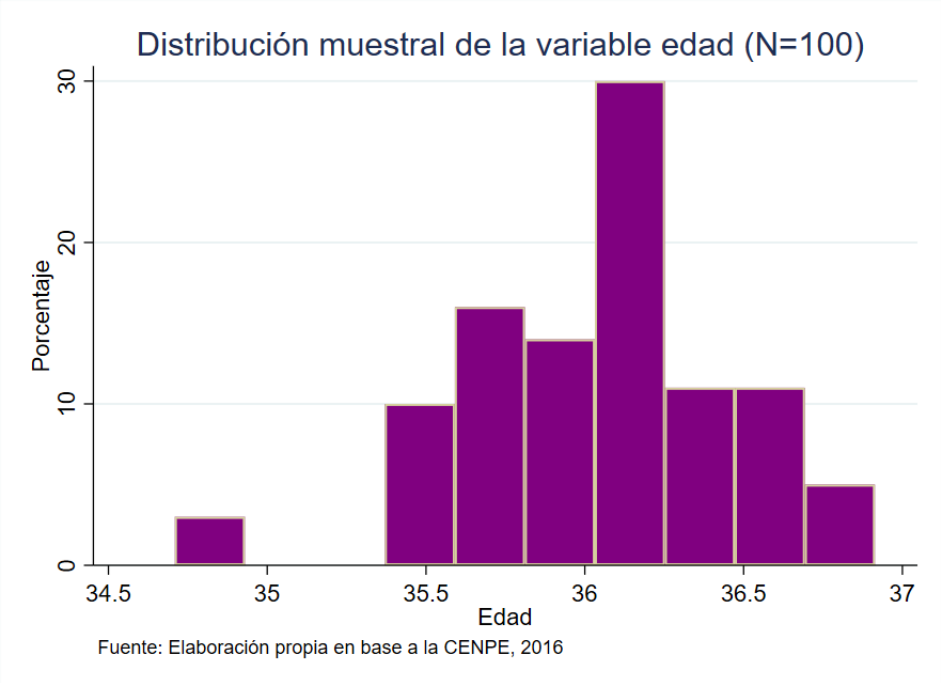


En el presente gráfico se puede observar que la distribución está sesgado hacia la derecha, por lo tanto no tiene un forma de una distribución normal. Asimismo, se observa que la media se encuentra entre 35 y 36.5.

Jarque-Bera test for Ho: normality

- Jarque-Bera normality test: 7.137
- Chi(2): 0.0282

Como el Chi(2) es menor al p-value ($0.02 < 0.05$), entonces la hipótesis nula de normalidad puede rechazarse con el 5% . De esta manera, se demuestra que con 10 iteraciones, la distribución muestral que se obtiene no es una normal. En el test de Skewness-Kurtosis, se obtiene un Chi(2) de 0.0054 y como es menor al p-value, entonces tambien se rechaza la hipótesis de normalidad.

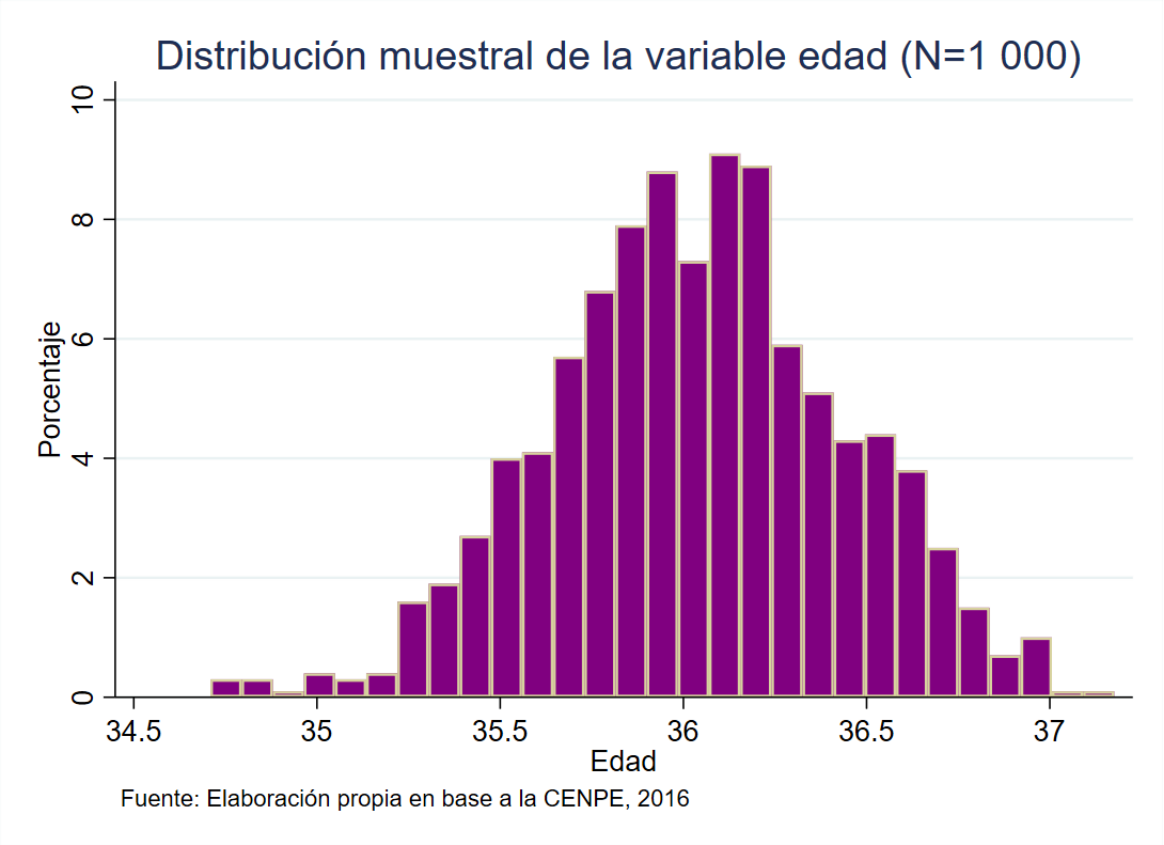


En el presente gráfico se puede observar que mientras la muestra aumenta de 10 a 100, la distribución comienza a tomar la forma de una normal. Además, su media está entre 35.5 y 37.

Jarque-Bera test for Ho: normality

- Jarque-Bera normality test: 10.71
- Chi(2): 0.0047

Como el Chi(2) es mayor al p-value ($0.004 < 0.05$), entonces la hipótesis nula de normalidad puede rechazarse con el 5 % de significancia. De esta manera, se demuestra que con 100 iteraciones, la distribución muestral que se obtiene no es una normal. Esto va en línea con el Chi(2) = 0.0122 obtenido del test de Skewness-Kurtosis, ya que también se rechaza la hipótesis nula de normalidad al 5 % de significancia.



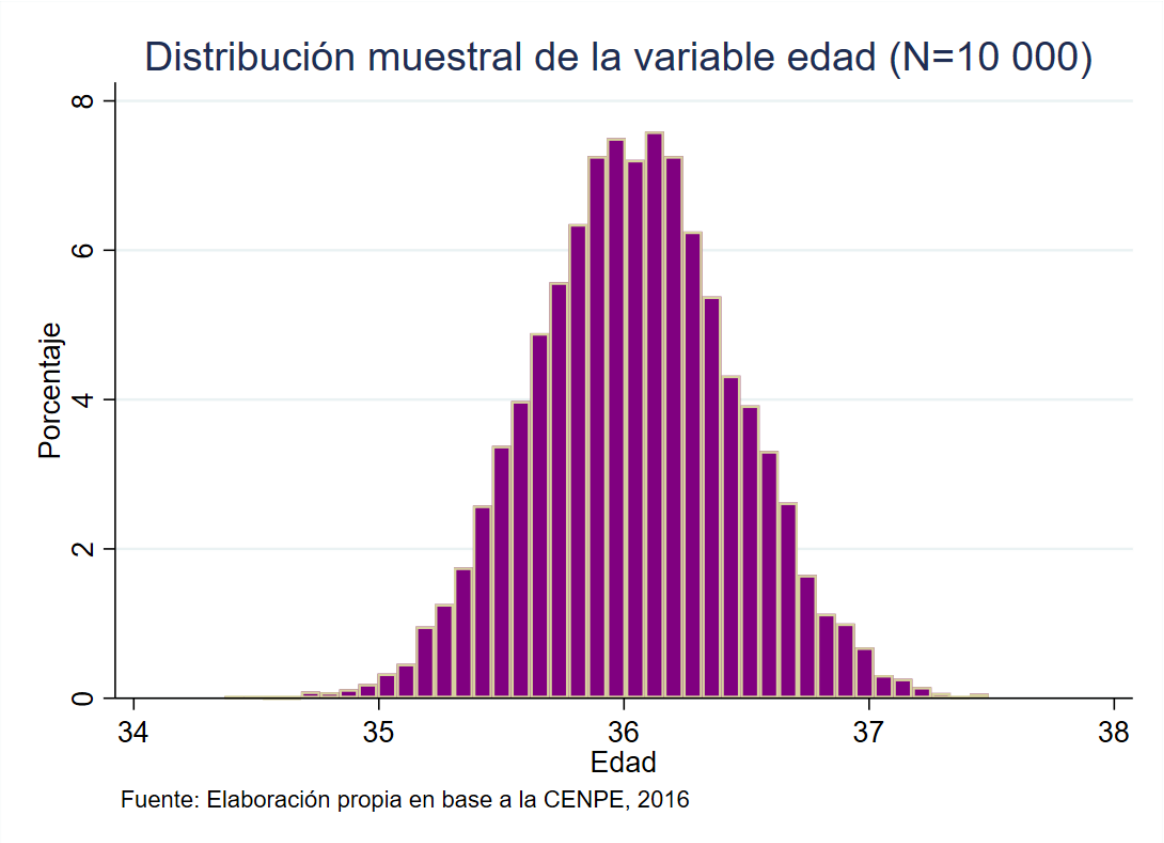
En la imagen previa, se observa que cuando el número de muestras tomadas es 1 000, entonces, la distribución muestral tiene la forma de una normal casi perfecta, con dos colas y con un concentración en la media.

Jarque-Bera test for Ho: normality

- Jarque-Bera normality test: 1.722
- Chi(2): 0.4228

Como el Chi(2) es mayor al p-value ($0.4228 > 0.05$), entonces la hipótesis nula de normalidad no puede ser rechazada con el 5 % de significancia. De esta manera, se demuestra que con 1 000 iteraciones, la distribución muestral de la variable edad se parece a una normal. Por otra parte, en el test de Skewness-Kurtosis con, se obtiene un Chi(2) = 0.4254 y, por ende, no se rechaza la

hipótesis nula de normalidad.



Por último, cuando se toman 10 mil muestras, la distribución muestral tiene la forma de una normal con media en 36 años y con un intervalo de 35 a 37 años de edad.

Jarque-Bera test for Ho: normality

- Jarque-Bera normality test: 0.3095
- Chi(2): 0.8566

Como el Chi(2) es mayor al p-value ($0.8566 > 0.05$), entonces la hipótesis nula de normalidad no puede ser rechazada con el 5 % de significancia. De esta manera, se demuestra que con 10 000 iteraciones, la distribución muestral de la variable edad es una normal. Asimismo, esto se

complementa con el Chi(2) = 0.8455 obtenido del test de Skewness-Kurtosis, ya que también no se rechaza la hipótesis nula de normalidad.

5. Parte 5 – Apropiación de conocimientos

Responder las siguientes preguntas sin utilizar jerga técnica y asumiendo que su oyente tiene conocimiento nulo de muestreo (prestar atención al límite de palabras de cada pregunta).

1. Un hacedor de políticas públicas sin conocimientos de estadística le pide que utilice la Encuesta Nacional Agropecuaria más reciente para obtener estadísticas distritales de acceso a asistencia técnica en agricultores de Ayacucho. Estos datos serán utilizados en el diseño de un programa público de apoyo a los productores. Explique en castellano sencillo y de forma intuitiva al funcionario si esto es posible y por qué. En caso no sea posible, propóngale qué otra fuente de datos podría usarse y cuáles serían sus ventajas y desventajas. (Palabras: 300)

   Respuesta:

   La Encuesta Nacional Agropecuaria permite construir indicadores del sector agropecuario, de manera que dicha información pueda orientar el diseño de políticas públicas que mejoren las condiciones de los agricultores peruanos. En el módulo de servicios de extensión agraria (Módulo 1543) se obtiene información sobre las capacitaciones y asistencia técnica. Por ende, la encuesta tiene la información necesaria para conocer si es que los productores recibieron asistencia técnica o no. Sin embargo, los niveles de inferencia de la ENA son: nacional, regional natural y departamental, por lo que no sería posible utilizar la ENA para construir indicadores a nivel distrital de la región de Ayacucho. En otras palabras, solo podríamos construir estadísticos confiables para el Perú, para las regiones naturales y para cada departamento, pero no podemos construir un estadístico confiable para. Por ejemplo, el distrito de Chilcas en Ayacucho. Sin embargo, se cuenta con información auxiliar como el Censo Agropecuario (CENAGRO) del 2012, es posible utilizar la metodología de “Small Area Estimation” que nos permita combinar la ENA y la CENA-

GRO con el objetivo de estimar el porcentaje de productores que reciben asistencia técnica de los distritos de Ayacucho sin la necesidad de generar una nueva fuente de información, es decir, sin realizar una encuesta que tenga la capacidad de representar a la población a nivel distrital. Aquello es posible, pues de acuerdo a Escobal, Fort y Zegarra (2015)[3] , las variables de la ENA y el CENAGRO son compatibles, lo cual mejora la capacidad de inferencia de los resultados que obtengamos.Considerando este punto, se puede concluir que si podemos utilizar la ENA para construir indicadores distritales de acceso a asistencia técnica, pero utilizando un modelo de estimación de áreas pequeñas.

2. En el Ministerio de Desarrollo le han encargado a Ud. como muestrista principal diseñar una encuesta probabilística para medir la prevalencia de vulnerabilidad a la pobreza (variable dicotómica) en cada región del Perú. Su supervisor, quien es un muy buen policymaker pero sin ninguna formación cuantitativa ni de muestreo, se enteró del libro de Dillman (2014) y de las tablas de cálculo de tamaño de muestra. Él está solicitándole que justifique sus decisiones metodológicas pues argumenta que "como máximo se necesitaría una muestra de tamaño n=1067 como se observa en la figura 3.5 de libro".

   Ud. ha planteado que se levante un muestreo complejo (estratificado y biétapico) para lograr el objetivo. Explique en castellano sencillo y de forma intuitiva si este tamaño de muestra satisfará sus requerimientos. Apóyese de los conceptos de tamaño de muestra efectivo, correlación intraclúster y efecto de diseño para su respuesta, pero en castellano simple. Recuerde que su supervisor no maneja estos temas. (Palabras: 500)

   Respuesta:

   Las tablas del libro de Dillman (2014) explica distintos tamaños de muestra para un muestreo aleatorio simple (MAS, en adelante. Sin embargo, el MAS no es el más apropiado si es que queremos medir la prevalencia de vulnerabilidad a la pobreza, debido a que puede resultar muy costoso y sobrepasarse del presupuesto asignado para obtener dicha infor-

[3]Escobal,J., Fort,R. y Zegarra, E. (2015). Agricultura peruana : nuevas miradas desde el Censo Agropecuario. Lima. GRADE

mación. En segundo lugar, se necesita medir esta variable para cada región del Perú, por lo que no podemos usar un muestreo aleatorio simple, ya que no podríamos asegurar la representatividad de las poblaciones pequeñas, como, por ejemplo, Tacna y Tumbes. Por ello, con lo mencionado anteriormente proponemos realizar un muestreo complejo (estratificado y biétapico), ya que nos permitirá abaratar costos, utilizar la información adicional disponible para tener estadísticos de mayor precisión y lograr representatividad de las regiones con menores poblaciones.

Respecto al tamaño de muestra que se ha planteado, n = 1067, consideramos que no nos permitiría cumplir con nuestros objetivos. En primer lugar, nuestra propuesta, implica agrupar a la población de conjuntos de 400 viviendas aproximadamente (conglomerados), para luego elegir una cantidad de ellos aleatoriamente (primera etapa) y luego, finalmente, elegir a las viviendas también aleatoriamente (segunda etapa). No obstante, los individuos de los conglomerados, en los que hemos dividido la población, pueden tener características similares por diversos factores, lo que puede generar que respondan de manera similar en las encuestas. Esta similitud entre los individuos dentro de un conglomerado es medida por la correlación intraclúster y, por ende, su existencia hace que el tamaño de nuestra muestra aumente. Por ejemplo, si la correlación intraclúster es 1, esto quiere decir que las personas son casi idénticas, por lo que cualquiera sea el número que entrevistems dentro de ese conglomerado, solo valdrá como si hubiéramos entrevistado a una persona. En segundo lgar, usualme como la correlación intracluster es positiva, es decir, los individuos se parecen, entonces no vamos a tener mucha variabilidad en la información que recolectemos lo que aumenta la imprecisión de nuestro estimador de prevalencia de la vulnerabilidad de la pobreza. Esto es lo que se conoce como efecto del diseño. En tercer lugar, si entrevistamos una cantidad “n” en un muestreo aleatorio estratificado y bietápico es como si hubiéramos entrevistado a una cantidad mucho menor “m”, donde m¡n (tamaño de muestra efectiva). Entonces, mientras más grande es la correlación intracluster, más grande será el efecto del diseño en la precisión del estimador, es decir, aumentará su varianza en comparación a un muestreo aleatorio simple. Luego, mientras más grande

sea el efecto de diseño, la muestra efectiva es más pequeña, es decir, como si hubieramos encuestado a menos personas.

Por ejemplo, si solo entrevistamos a 1067 personas utilizando el muestreo estratificado y biétapico, las 1067 personas entrevistadas no valdrán 1067, sino que sería como si hubiéramos entrevistado a un grupo mucho menor. Además, la cantidad de 1067 sería solo representativa a nivel nacional, por lo que el tamaño de muestra que necesitaríamos para el muestreo estratificado y biétapico es mucho mayor.