FACULTAD DE



**CIENCIAS SOCIALES**

**Trabajo 1 – Teoría de Muestreo**

Indicaciones generales:

* Este trabajo se realizará en **grupos de** **cuatro estudiantes**. Los grupos serán armados por el profesor utilizando las parejas formadas para la Tarea 1.
* Se entregará un solo archivo que contenga los códigos utilizados, resultados y respuestas escritas (cuando sea necesario). Para esto puede utilizarse un archivo Markdown (R) o un Jupyter Notebook (enlazado a Stata). Estos archivos deben estar ordenados y comentados. Los comentarios deben indicar claramente cada sub-sección del Trabajo y explicar los pasos seguidos. Junto con el trabajo se entregará el contrato de trabajo grupal que se encuentra al final del documento. Este debe ser firmado por todos los miembros del grupo.
* Se valorará en la calificación el orden y claridad del código. Este debe ser replicable y llegar a los mismos resultados que se reporten como respuestas a las preguntas.
* Para extraer muestras y declarar el diseño muestral, utilizar los comandos utilizados en las PDs (sample/slice\_sample y svydesign en R, sample y svyset en Stata)
* La fecha de entrega será el **domingo 03/08 hasta las 14:00 horas (hasta las 2 de la tarde del domingo)**. A las entregas que se realicen tarde se les descontarán dos puntos por cada hora transcurrida desde la hora de entrega. Luego de la tercera hora de retraso, ya no se recibirá el trabajo y se calificará con cero.

Indicaciones del Trabajo 1

**Parte I (7 puntos)**

1. Descargue el Censo Nacional de Población Penitenciaria 2016 de la página del INEI (<http://iinei.inei.gob.pe/microdatos/>). Puede descargar solo la carátula o también más módulos de dicho censo.
2. Genere una variable dicotómica con el valor de 1 para quienes cometen “Delitos contra la libertad”, 0 para los demás casos. Llamar a esta variable “pp\_dcl”
3. Establecer una semilla (“seed”) usando los dígitos del código PUCP de alguno de los miembros del grupo.
4. Extraer una muestra estratificada y por conglomerados que contenga entre 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.
5. Declarar el diseño muestral (completo) y obtener las siguientes estadísticas (deben ser representativas):
   1. Porcentaje de personas que ha cometido “Delitos contra la libertad”
   2. Total de personas que ha cometido “Delitos contra la libertad”

**Parte II**

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). Para esto se recomienda revisar la ficha técnica de la encuesta. Esta se encuentra disponible en la web de microdatos del INEI. (2 puntos)
2. Elegir y calcular dos variables: (i) una dicotómica, (ii) una continua. Por lo menos la variable dicotómica debe ser una variable **definida y creada por ustedes**. (2 puntos)
3. Para la dicotómica, obtener la media y población. Para la continua, obtener la media (deben ser representativas). (3 puntos)

**Parte III (6 puntos)**

Responder uno de los siguientes bloques de preguntas (elija entre Bloque A o Bloque B). Prestar atención al límite de palabras de cada pregunta.

**BLOQUE A**

Pregunta 1 (Palabras: 300)

Suponga que usted trabaja en el Área de Estudios del Ministerio de Educación. En su área quieren medir el nivel de bienestar psicológico en estudiantes de educación secundaria en Lima. Para ello, el psicólogo especialista sugiere utilizar como instrumento la Escala de Bienestar Psicológico de Ryff. A usted le piden encargarse del diseño muestral de la encuesta a levantar. Explique cuál sería el marco muestral que usaría y por qué. Comente la calidad del marco muestral elegido (incluya los conceptos de población objetivo, población encuestada). Además, explique cuál sería el diseño muestral que usted sugeriría para la encuesta y por qué. Comente, de manera breve, las ventajas y desventajas del diseño elegido frente a los otros. Finalmente, de manera sencilla e intuitiva, comente el tamaño de muestra que elegiría (n) y a qué factores responde el tamaño propuesto.

Pregunta 2 (Palabras: 500)

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 todas las regiones 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 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** por qué necesitaría (o no) una muestra de mayor tamaño. Incluya en su respuesta el concepto de tamaño de muestra efectivo, correlación intracluster y efecto de diseño. Recuerde que su supervisor no maneja estos temas.

**BLOQUE B**

Pregunta 1 (Palabras: 500)

Suponga que usted es ascendido como muestrista principal en el Ministerio de Salud. Su primera tarea es terminar de definir el diseño de una encuesta probabilística para medir el nivel de desnutrición en niños de educación primaria. El anterior muestrista había sugerido un diseño por conglomerados donde los conglomerados sean los distritos. Discuta si esta opción le parece la más idónea. ¿Qué beneficios y/o perjuicios traería hacerlo de dicho modo? Además, si usted considera que hay otra opción, explique cuál sería y por qué. Incluya los conceptos de UPM y USC.

Un muestreo por conglomerados de una sola etapa resultaría muy costoso y poco eficiente. Por un lado, si se randomiza la selección de distritos a nivel nacional para luego “censar” a los niños en educación primaria puede llegar a ser muy costoso porque no necesariamente todos los hogares del distrito cuentan con niños en educación primaria. Es decir, tendrías que censar para asegurar que estas consiguiendo la información que necesitas. Por otro lado, al censar en dentro de los distritos seleccionados aleatoriamente es probable que exista una alta correlación intraclusters (es probable que los niños del mismo distrito cuenten con características socioeconómicas homogéneas) lo cual ofrece poca varianza sobre la variable de interés lo cual implica pérdida de eficiencia significativa en el muestreo por conglomerados respecto a un muestreo aleatorio simple (el efecto de diseño o *deff* sería mayor a 1).

Creemos que se puede establecer un mejor diseño muestral para medir indicadores de desnutrición en la población de escolares de educación primaria. Para ello, una primera observación propuesta es delimitar mejor la población de interés. Se propone delimitar la muestra a alumnos del primer, tercero y sexto grado de primaria con la finalidad de poder medir evaluar el indicador de interés al inicio de la primaria, a mitad del nivel escolar y a punto de culminar el mismo. Una vez delimitado el universo de interés se propone pasar a un muestreo por conglomerados en dos etapas. Se define como unidad primaria de muestreo (UPM) a las Unidades de Gestión Educativa Local (UGEL) con colegios de nivel primario. Estas agrupan colegios incluso de distintos distritos y abarcan todo el territorio nacional. Luego, se define como unidad secundaria de muestreo a colegios que ofrecen servicios educativos de nivel primario dentro de cada UGEL. Una vez seleccionados los colegios como parte de la muestra final, se recogerá información estilo “censo” de cada uno de los alumnos de que conforman el primer, tercero y sexto grado de primaria.

Suponga, además, que su superior le dice que han aumentado el presupuesto para el estudio, por lo cual pueden incrementar el tamaño de muestra calculado por el anterior muestrista. En ese sentido, le consulta si resulta más conveniente aumentar los conglomerados o más bien el número de estudiantes dentro de cada conglomerado. Argumente su respuesta.

Se podría evaluar la posibilidad de añadir una tercera unidad de muestreo dentro de cada grado para seleccionar solo un grupo de niños y niñas. Además, se puede aprovechar esta tercera etapa para estratificar la muestra según sexo del escolar. Creemos que esta tercera etapa estratificada puede disminuir el valor del deff dado que el indicador de desnutrición no en los escolares no estaría fuertemente ligado a la selección de la UGEL ni de la escuela, lo cual nos permite ganar eficiencia con el diseño

Discuta las ventajas y/o desventajas del diseño planteado por al anterior muestrista (diseño por conglomerados) frente a un diseño por estratos. De optar por un diseño por estratos, cuáles serían las ventajas y desventajas frente al diseño inicialmente planteado. Incluya conceptos como homogeneidad, heterogeneidad, efecto de diseño.

El diseño planteado por el muestrista favorece la reducción de costos de implementación de la encuesta. Sin embargo, sabemos que el efecto de diseño puede ser superior a 1 (lo cual implica una pérdida de eficiencia respecto a un muestreo aleatorio simple) si es que la correlación intracluster es muy alta. Ello debido a que al hacer un muestreo por conglomerados se espera que las primeras unidades de muestreo sean homogéneas entre los grupos y heterogéneos entre las segundas unidades de muestreo y al no ser los colegios diferentes entre sí, puede perjudicar el efecto de diseño. En cambio, un muestreo estratificado reduce el efecto de diseño, sin embargo, es mucho más costoso ya que consideraría a toda la población.

Pregunta 2 (Palabras: 300)

Un hacedor de políticas públicas que cuenta con conocimientos intermedios 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. Explíquele 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.

La ENA más reciente es la del 2022. De acuerdo con la ficha técnica de esta encuesta, no es posible obtener información **distrital** confiable respecto al acceso a asistencia técnica de agricultores en los distritos de Ayacucho. Esto debido a que los niveles de inferencia o los dominios que fueron utilizados en el diseño muestral sólo permiten obtener indicadores confiables hasta el nivel departamental para pequeños y medianos productores agropecuarios. Esto implica que no se obtuvo una muestra aleatoria para cada distrito, lo cual nos permitiría obtener estimadores insesgados a nivel distrital.

De hecho, como se menciona en la documentación de la encuesta “La muestra fue probabilística, estratificada, bietápica, independiente en cada departamento, de selección sistemática con probabilidad proporcional al tamaño en la primera etapa y sistemática simple en la segunda etapa.”. Esto implica que se desarrolló un muestreo complejo independiente para cada departamento y, dentro de cada uno, se estratificó por región natural y tamaño de la unidad agropecuaria, y se definieron 2096 conglomerados los cuales fueron aleatorizados en una primera etapa con pesos heterogéneos y luego dentro de cada conglomerado se realizó una aleatorización simple.

De lo anterior, se puede reafirmar que no es lo óptimo obtener estadísticas distritales de la ENA porque: (i) puede que los distritos no hayan sido seleccionados en la primera etapa (aleatorización de conglomerados), (ii) puede que existan algunas unidades agropecuarias, pero estas no forman parte de una muestra aleatoria del distrito por lo cual no garantiza la insesgadez de las estimaciones que se puedan realizar.

Lo recomendable para poder hacer las estadísticas que propone el hacedor de políticas públicas es utilizar el Censo Nacional Agropecuario que, aunque es del 2012, sí permite obtener la información requerida a nivel distrital de manera certera puesto que no es una muestra, sino un censo.

**Contrato de grupo de trabajo**

**Nombres completos de cada miembro:**

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. Valeria Lévano Torres

**Objetivos y fechas**

Rellenar con las tareas que consideren necesarias para completar el trabajo grupal. Asignar un encargado/a, fecha de entrega y si se cumplió con la tarea asignada.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tareas** | **Encargado/a** | **Fecha de entrega** | **¿Se cumplió?** |
| 1. | Cesar | 02/09/23 | si |
| 2. | Mauricio | 02/09/23 | si |
| 3. | Andrea | 02/09/23 | si |
| 4 Principalmente la parte II | Valeria | 02/09/23 | si |
| Añadir según sea necesario |  |  |  |

**Sobre la relación entre miembros del grupo**

Normas grupales: Consideramos importantes las siguientes actitudes y comportamientos y nos esforzaremos en mantenerlas a lo largo del trabajo grupal.

|  |
| --- |
| Coordinar y revisar el trabajo completo. |

Toma de decisiones: Tomaremos decisiones de la siguiente manera:

|  |
| --- |
| Comunicarnos ante eventuales dudas y si es necesario con profesores. |

**Reglas de comunicación**

Los siguientes medios serán utilizados para coordinar los aspectos relacionados al trabajo grupal. Si el grupo tendrá reuniones de grupo (presenciales o virtuales), especificar las fechas de reunión.

|  |
| --- |
| Reuniones virtuales, mensajes. |

**Resolución de conflictos**

Especificar cómo el grupo lidiará con los conflictos que pueda haber o situaciones en los que no se cumpla este contrato.

|  |
| --- |
| Comunicación primero con el grupo y después con profesores. |

**Firmas**

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. Valeria Lévano Torres

Al escribir nuestros nombres completos, aceptamos la responsabilidad de completar las tareas del trabajo grupal.