 CONTROL Muestreo – 12/octubre - Entrega 14/octubre

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

Facultad de Matemáticas, Departamento de Estadística

**DIPLOMADO EN ESTADISTICA – VERSION 2023 Online**

Nombre: Martín Venegas Márquez

Fecha de nacimiento: 6 de mayo del 97 (06051997)

\*NOTA ESTUDIANTE 1\*: Se utilizará color azul para diferenciar las partes del control que corresponden a la respuesta del estudiante.

\*NOTA ESTUDIANTE 2\*: Cabe destacar que las estimaciones realizadas para este control, así como las interpretaciones asociadas asumen que el conjunto de datos entregados (Salud2017.xlsx) corresponden a un marco muestral de individuos. De este marco muestral, se realizan cálculos de muestra para realizar una encuesta de salud hipotética y estimaciones de parámetros asumiendo que ya se hizo el trabajo de recolección de esa encuesta hipotética.

El MINSAL liberó una base de datos con 6233 casos analizados en el contexto de salud. Entre la información disponible destaca el IMC, sexo, edad y la presencia de enfermedades tales como Hipertensión, Diabetes y Asma, entre otras. Sobre esta información usted, junto a su(s) colega(s), tendrán que trabajar usando la librería ***samplingbook*** (datos - ver base SALUD2017.XLSX).

**Primera parte:** Interesa estimar, con un 95% de confianza:

1. Proporción de diabéticos con un error del 5%
2. El nivel medio del IMC con un error de un punto.

Para la determinación óptima del tamaño muestral utilice la siguiente información disponible de la encuesta nacional de salud previa: El 12% padece de diabetes y el IMC medio fue de 29 con una desviación estándar de 5 puntos.

Use como semilla en la selección **set.seed(fecha nacimiento del jefe de grupo)** – la que se debe indicar en el informe. Proceda a seleccionar desde la base SALUD2017 el número de casos determinado en el punto anterior, y con esta información complete la siguiente tabla:

**En resumen, obtenga una m.a.s. de tamaño n y “rellene” la siguiente tabla:**

**Tabla 1. Tamaños de muestra y estimaciones para encuesta de salud hipotética con MAS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Grupo**  **Etario** | **Número de Casos** | | **IMC medio** | | **Proporción de Diabéticos** | |
| **Base** | **Muestra** | **Estimación** | **ErrStandard** | **Estimación** | **ErrStandard** |
| 15-29 | **1296** | 27 | 26.6 | 5.34 | .037 | .037 |
| 30-59 | **3010** | 82 | 29 | 4.94 | .085 | .031 |
| 60 y más | **1927** | 59 | 29.7 | 6.23 | .237 | .055 |
| **TOTAL** | **6233** | 168 | 28.8 | .422 | .131 | .025 |

**\*\* sin utilizar strata, sino que el método clásico de media y proporción (Smean y Sprop).**

En la Tabla 1 se presentan los tamaños de muestra y las estimaciones solicitadas para nuestra encuesta de salud hipotética. En concreto, se estima el IMC medio con un error de un punto y un antecedente de desviación estándar de 5 puntos, así como la proporción de diabéticos con un error del 5% y una proporción de antecedente del 12%. Ambas estimaciones se realizan considerando el total de individuos, así como también separando por grupo etario.

En base a la información anterior, se aplica un muestreo aleatorio simple (MAS), el cual sugiere un tamaño de muestra mínimo de 95 casos para la proporción de diabéticos y 159 casos para la estimación del IMC medio. Siguiendo un criterio maximalista, sería prudente escoger una muestra de 159 casos, sin embargo, también es necesario considerar un porcentaje de sobre muestreo, ante una potencial no respuesta de parte de los informantes. Tomando como antecedente las tasas de respuesta (78.7%) de la Encuesta de Calidad de Vida y Salud 2015-2016 (ENCAVI), se establece como porcentaje de sobre muestreo la tasa de no respuesta de la ENCAVI, es decir, el 21.3% de 159. **El tamaño de muestra considerando el sobre muestreo es de 193 individuos.**

Una vez obtenida la muestra, se observa que, de los 193 casos encuestados, 25 casos no tienen un dato valido para IMC. Se excluyen estas unidades de la muestra, lo cual da por resultado una muestra de 168 casos, la cual está por sobre el tamaño muestral mínimo (n = 159). **Se usará esta muestra de 168 casos válidos para las estimaciones.**

La estimación total para el IMC medio es de 28.8 con un error estándar de .422. Este error es más bajo que el previsto en un principio, ya que la muestra utilizada para los cálculos fue ligeramente más grande (+ 9 casos). Al revisar las estimaciones por grupo etario, vemos que el IMC medio es de 26.6 para el grupo de 15 a 29 años, 29 para el grupo de 30 a 59 años y 29.7 para el grupo de 60 y más años. Cabe destacar que los errores estándares de las estimaciones por grupo erario son considerablemente grandes, entre 5 y 6 puntos aproximadamente. Esto sugiere que las estimaciones dentro de cada grupo etario son menos precisas.

En el caso de la proporción de diabéticos, la estimación total es de 13.1% con un error estándar de 2.5%, el cual también es menor que el originalmente considerado, principalmente por el tamaño de muestra. Las estimaciones por grupo etario son: 3.7% para el grupo joven, 8.5% en adultos y 23.7% en adultos mayores. En lo que respecta a los errores estándar, vemos que están entre el 3% y el 5%, lo cual está en línea con la precisión requerida originalmente.

**Segunda parte**

Es común en el área de salud, presentar resultados estratificados de acuerdo con el grupo Etario (categorías de edad). De acuerdo con el último informe del Minsal, se presentan resultados según tres categorías de edad y se entregan los resultados observados en la encuesta previa:

**Tabla 2. Antecedentes ENS anterior**

|  |  |  |  |
| --- | --- | --- | --- |
| Grupo etario | % poblacional | % de Diabéticos | IMC (media +/- desv.est.) |
| 15-29 | 20% | 4% | 24 +/- 4 |
| 30-59 | 50% | 10% | 28 +/- 5 |
| 60 y más | 30% | 24% | 34 +/- 7 |

Con base a esta información (histórica) determine el tamaño de muestra **óptimo** (no proporcional) para estimar:

* Proporción de diabéticos con un error del 5%.
* IMC medio con un error no mayor a 0.7 puntos.

*Ayuda: utilice las proporciones conocidas 4%, 10% y 24%, como información para obtener los Sh =sqrt(Ph(1-Ph)) a ser incorporada en la fórmula del* ***stratasize y stratasamp****.*

Obtenga la muestra estratificada óptima y usando **nuevamente la semilla** y “**rellene” la siguiente tabla:**

**Tabla 3. Tamaños de muestra y estimaciones para encuesta de salud hipotética con ME**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Grupo**  **Etario** | **Número de Casos** | | **IMC medio** | | **Proporción de Diabéticos** | |
| **Base** | **Muestra** | **Estimación** | **ErrStandard** | **Estimación** | **ErrStandard** |
| 15-29 | **1296** | 49 | 26.33 | .66 | 0 | NA |
| 30-59 | **3010** | 123 | 30.01 | .50 | .186 | .034 |
| 60 y más | **1927** | 76 | 29.39 | .61 | .302 | .051 |
| **TOTAL** | **6233** | 248 | 29.05 | .33 | .183 | NA |

En la Tabla 3 se presentan los tamaños de muestra y las estimaciones solicitadas para nuestra encuesta de salud hipotética, pero ahora estratificando por grupo etario. Al igual que en el ejercicio anterior, se estima el IMC y la proporción de diabéticos, pero ahora considerando un 0.7 puntos y 5% de error, respectivamente. Se utilizan los antecedentes de la Tabla 2 para los cálculos de tamaño de muestra.

Se emplea un muestreo estratificado (ME) con afijación óptima. Para la proporción de diabéticos, los cálculos de muestra sugieren un n de 152 casos (15-29 años: 32 casos; 30-59 años: 73 casos; 60+ años: 47 casos). En cambio, para la estimación del IMC medio, los cálculos de muestra sugieren un n de 222 casos (15-29 años: 46 casos; 30-59 años: 107 casos; 60+ años: 69 casos). Se toma como punto de partida los n sugeridos para la estimación del IMC medio y se le suma un porcentaje de sobre muestra por cada estrato, basándose en la tasa de no respuesta de la ENCAVI 2015-2016 (23.1%). Es decir, se agrega un 23.1% del n sugerido para cada estrato, la cual da por resultado los siguientes tamaños de muestra por estrato: 15-29 años: 57 casos; 30-59 años: 132 casos; 60+ años: 85 casos. Esto da un tamaño de muestra total con sobre muestro de 274 casos.

Posterior a “aplicar la encuesta”, se evalúan los casos perdidos en el IMC y se filtran las unidades que no tengan una respuesta valida en esta variable. **Este filtro da por resultado una muestra de 248 casos, con la siguiente distribución:** **15-29 años: 49 casos; 30-59 años: 123 casos; 60+ años: 76 casos.** Estos tamaños de muestra por estrato cumplen el n mínimo sugerido por los cálculos de muestra para el IMC medio con la precisión requerida, por lo que se emplearán para realizar las estimaciones.

La estimación del IMC medio da 29.05 para el total y 26.33, 30.01 y 29.39 para los grupos de jóvenes, adultos y adultos mayores, respectivamente. Los errores estándar están dentro de la precisión requerida (< .07 puntos). De esto se desprende que la estratificación por grupo etario entrega estimaciones mucho más precisas que las estimaciones realizadas con MAS, las cuales estaban por sobre los 5 puntos de error.

En el caso de la estimación de la proporción de diabéticos, se estima un 18.3% en total, 18.6 para el grupo adulto y 30.2% para el grupo de adultos mayores. En lo que respecta al grupo jóvenes, la estimación da un 0%, esto ya que, en la muestra recopilada, ningún joven resultó ser diabético. Esto afecta el calculo de los errores estándar del grupo de jóvenes, así como del total. Los errores estándares que si pudieron ser calculados (adultos y adultos mayores) están en línea con la precisión requerida.

**DESAFIO - resolver UNO de los 3 desafíos.**

**D**iscutan y escriban SUCINTAMENTE (no más de una página por desafío) la forma de llevar a cabo. Es decir, debe responder cada punto que se presenta a continuación:

* ¿Cuál es la unidad de información?
* ¿Cuál es la unidad muestral? (si no coinciden, explique el por qué)
* ¿Hay “agrupaciones” o conglomerados naturales, alguna estratificación? ¿cuál o cuáles?
* **¿Cómo construye u obtiene el marco muestral? (trate de ser lo más explícito)**
* ¿Qué características, idealmente, como mínimo debe poseer el marco muestral?
* Si dispone de presupuesto para realizar encuestas propuestas, ¿qué tipo de muestreo aplica? ¿cómo los distribuye los casos? ¿qué criterios utilizaría?

1. Producto de la pandemia, la autoridad desea estimar la proporción de niñas y niños con edades entre 2 y 6 años que no concurren al jardín escolar, cuyas madres no cuenten con educación terciaria (técnica, profesional), que estén interesadas en enfrentar desafíos laborales o de emprendimiento. El estudio se requiere para la región del Maule y debe considerar las comunas con más de 40 mil hab. Se deben realizar 1.200 encuestas (aprox) en el área urbana (ideal – presencial).

**\*\* no se dispone del registro social de hogares (o equivalente)**

1. Se desea construir un indicador sobre la “seguridad de tránsito” en una comuna (o ciudad), para la evaluación de los puntos muestrales se dispone de un conjunto de alumnos de ingeniería en transporte y tránsito, quienes podrán visitar durante el día los puntos y evaluarlos (se calcula que podrán evaluar cerca de 600 puntos).
2. Una compañía distribuidora de gas natural de una región sureña, a instancia de la superintendencia de Gas y Combustibles, debe entregar una estimación de la materialidad de la superficie por la cual se encuentra las tuberías de gas, cuya extensión es de poco más de 820 kms. Para lo anterior, la compañía nos envía una base de datos con la siguiente información:

**Cod\_comuna**: código de la comuna en la que se encuentra el tramo de tubería.

**Fecha\_inst**: fecha de instalación en formato dd-mm-aaaa.

**Shape\_len**: largo – en metros – del tramo de tubería.

**Identifica**: código que permite identificar en forma única al tramo de tubería.

De acuerdo con los costos involucrados, la empresa dispone de equipos para evaluar la materialidad en poco más de 5.000 puntos.

**Respuesta desafío 1)**

¿Cuál es la unidad de información?

La unidad de información refiere a la fuente de donde se obtendrá la información. En este caso, corresponde a los informantes que residan en las comunas urbanas de más de 40 mil habitantes de la región de Talca, en tanto serán ellos quienes entreguen información de la edad de los niños, su asistencia al jardín e información de la madre.

¿Cuál es la unidad muestral? (si no coinciden, explique el por qué)

La unidad muestral serán las manzanas. La unidad muestral está estrechamente ligada a la disponibilidad de marco muestral y el presupuesto asignado al estudio. El escenario ideal sería contar con un registro a nivel de individuo (por ejemplo, RSH), sin embargo, a falta de este la unidad más desagregada con la que se puede contar haciendo uso de datos públicos (Censo 2017) serían las manzanas. Es decir, la unidad de información y la unidad muestral no coinciden por la disponibilidad del marco muestral.

¿Hay “agrupaciones” o conglomerados naturales, alguna estratificación? ¿cuál o cuáles?

Independiente de la disponibilidad del marco muestral, los conglomerados naturales son las comunas en tanto agrupan manzanas, y las manzanas en tanto agrupan viviendas, las viviendas en tanto agrupan hogares y los hogares en tanto agrupan individuos. Si se considera el marco muestral que se ha señalado hasta ahora, solo se podría emplear la comuna y manzana para un muestreo de conglomerados, las unidades más desagregadas tendrían que ser censadas.

En lo que respecta a la estratificación, sería ideal poder estratificar por alguna característica necesaria para hacer la estimación solicitada (p. ej nivel educacional de la madre o edad de los niños). Sin embargo, con el marco muestral que se ha señalado hasta ahora, sería más prudente estratificar por comuna, con tal de asegurar que las estimaciones a nivel comunal sean lo suficientemente precisas.

**¿Cómo construye u obtiene el marco muestral? (trate de ser lo más explícito)**

La primera tarea sería utilizar la información geográfica que entrega públicamente el Censo 2017. Con esta información, se podrían tomar dos opciones, dependiendo del presupuesto con el que se cuente. La primera opción sería utilizar solamente la información disponible, y hacer un muestreo tomando como unidad más desagregada las manzanas. Esto conllevaría muestrear manzanas y encuestar todas las viviendas dentro de cada manzana elegida aleatoriamente. La segunda opción sería intentar tomar la información de todas las manzanas de las comunas de más de 40 mil habitantes y construir un marco muestral de viviendas. Con ese marco muestral se podría hacer un muestreo de viviendas y encuestar a un informante idóneo para cada vivienda. Sin embargo, en un escenario realista, esta última opción es poco viable ya que la construcción de un marco muestral es una gran empresa en si misma y utilizaría una enorme cantidad de recursos.

¿Qué características, idealmente, como mínimo debe poseer el marco muestral?

La función de un marco muestral es tener una referencia de la población con información que ayude a efectuar un muestreo. Si bien la información necesaria puede variar de acuerdo al tipo de muestreo (p. ej si se quiere estratificar por NSE de la vivienda, se debe contar con esa información en el marco), la información mínima necesaria que debe poseer un marco muestral es una forma de localizar geográficamente a la unidad de información.

Si dispone de presupuesto para realizar encuestas propuestas, ¿qué tipo de muestreo aplica?

¿cómo los distribuye los casos? ¿qué criterios utilizaría?

Asumiendo que se cuenta con el presupuesto suficiente, consideraría un muestreo de manzanas, estratificado por comunas tomando como población de referencia las comunas de más de 40 mil habitantes de la región de Talca (que según el Censo 2017 son 10 comunas). Además. Consideraría una afijación proporcional, con tal de no sub o sobre representar a alguna comuna. De esta manera, las estimaciones serán precisas tanto a nivel de la región de Talca (urbana y en comunas +40 mil habitantes) y comunal.

Indicaciones:

* Redacte un informe (pregunta/respuesta) que permita responder cada punto indicado.
* NO coloque el script en el informe (solo tablas, gráficos, métricas de interés, valores resultantes que ayuden a respaldar sus respuestas, no relleno).
* Por ningún motivo el último taller es pauta para este control, recaude toda la información pertinente (teórica como práctica).
* El trabajo es individual, pero puede ser grupal también (hasta 3 personas, sin más excepciones).
* El informe completo se entrega en el buzón “Entrega Final” (disponible en la plataforma del curso). No olvide poner el nombre de los miembros del grupo en el encabezado del informe!.
* **Fecha de entrega: 14 de octubre hasta las 18:00 hrs.**
* Se debe entregar si o si un avance el 12 de octubre hasta las 21:30 hrs.