MUESTREO EN POBLACIONES FINITAS 2016

Fundamentos y Métodos

Trabajo Práctico 1
PARTE A

En el DataFrame "*Marco.PO.RData*" que acompaña al TP1 se dispone de información del Personal Ocupado (PO) correspondiente al año 2014 de un universo N=3984 unidades económicas de un sector industrial, actualizado por altas y bajas por diversas fuentes (registros administrativos, listados de las cámaras que las agrupan, etc).

El objetivo es diseñar una muestra estratificada para una encuesta continua de empresas que se repetirá cada 6 meses. La misma brindará estimaciones de distintos indicadores económicos que permitirán evaluar la demanda laboral, disponibilidad de puestos, las ventas, valores de producción, compras e insumos de materia prima, etc. La necesidad de la encuesta en parte se sustenta en que el periodo intercensal es de 5 años, un periodo muy largo para estudiar la dinámica del sector y se consideró que una encuesta semestral satisface estos requerimientos.

Para los fines del muestreo se consideró que PO es una variable estable en términos económicos y es la que sufrió menos modificaciones entre la fecha de los datos y la de diseño. Por otro lado las características principales a estudiar tienen un nivel de correlación aceptable (constatada por los cálculos sobre los datos del último censo) con la que las estimaciones se verán beneficiadas por su empleo como característica para estratificar.

Como paso previo al diseño se realizó un pequeño estudio en donde se seleccionaron previamente 50 empresas y se las indagó sobre algunas de las principales características a estimar por la futura encuesta. Este estudio o prueba piloto permitió ajustar distintos aspectos del operativo final: la evaluación de las preguntas más sensibles del cuestionario, el protocolo de ingreso a la empresa, el tiempo de ejecución de la encuesta, los distintos tipos de no respuesta, la performance de los encuestadores y supervisores, los programas de carga y estimación, los de imputación para datos faltantes, entre otros.

Del conjunto de variables investigadas en la prueba, VMP (Valor de la Materia Prima comprada en los últimos 6 meses) es una de las que se considera relevante para la encuesta. Del comportamiento de VMP y su correlación con PO permite sospechar que no se puede emplear directamente a PO como estratificadora y que hay que estudiar la discrepancia entre una y otra. El archivo "Pueba.Piloto.RData" detalla el valor que toman tanto PO (variable X) como VMP (variable Y) para las 50 empresas de la prueba piloto. La idea es asumir un modelo lineal simple de discrepancia entre PO y VMP para el algoritmo de Kozak, que fijará la estratificación definitiva incluido en el Package "stratification". Con la ayuda del paquete y proponiendo las siguientes opciones (1):

- CV=2%,
- Asignación óptima de Neyman dentro de los estratos,
- Un estrato de Autorepresentados,
- Tasa de respuesta del orden del 80% en cada estrato, salvo en el de Autorepresentados, en cuyo caso y para este estrato en particular se asumirá respuesta total en todas las evaluaciones,

se pide:

1. Constatar la asimetría de la variable PO y presentar evidencia de la misma. Evaluar la información provista por "Prueba.Piloto.RData" presentado un gráfico de PO vs VMP y proponiendo un modelo simple de regresión entre PO y VMP atendiendo la probable heterocedasticidad en el modelo. Con fines descriptivos acompañe con alguno de los plots que sugieren ésta heterocedasticidad (por ejemplo entre los residuos al cuadrado de la regresión y la variable PO u otros que encuentre adecuados a tal fin) o a través del resultado de un test (por ejemplo, el test de Breusch & Pagan u otro de su conocimiento) [1].

2. Aceptando la evidencia del punto 1 estimar según los datos provistos los parámetros "beta", "sig2" y "gamma" necesarios para atender el modelo lineal simple de discrepancia en el algoritmo de Kozak. Recordar que "stratification" permite modelar $Y_i = \beta X_i + \varepsilon_i$ con $\varepsilon_i \sim N\left(0,\sigma_i^2\right)$ y $\sigma_i^2 = \sigma^2 X_i^\gamma$ donde beta=β, sig2=σ^2 y gamma=γ son parámetros que deben ser brindados por el usuario [2].

3. Con los valores estimados del punto 2 más las opciones (1) estudiar las alternativas de 2 hasta 5 estratos con respecto al tamaño de muestra final según la precisión deseada. Presente una tabla resumen con las distintas alternativas y con la siguiente información: opciones empleadas en el algoritmo, los bordes de los estratos surgidos del algoritmo, N_h y n_h resultantes, los CV alcanzados en los estratos para la variable y el tamaño final de la muestra.

4. ¿Cuál de las soluciones estudiadas aconsejaría si se pretende adoptar aquella que determina el menor tamaño muestral que satisface los requisitos de precisión?

5. Asigne cada empresa del universo respetando la estratificación que propuso en el punto 4 y seleccione la muestra según los tamaños por estrato definidos por el algoritmo[3]. Acompañar la presentación del TP con la muestra seleccionada.

6. Si se desea después de un año renovar el 20% de la muestra en los estratos donde no se censa, reteniendo parte de la muestra original y seleccionando a las nuevas unidades por MSA en cada estrato de $U_h - s_h$, con h=1,…,H-1 (el H es el *takeall*) y s_h la muestra original en el estrato en cuestión. Proponer 2 alternativas para estimar la diferencia inter-anual $\hat{t}_{yt} - \hat{t}_{y(t-1)}$ entre dos totales de una misma característica. ¿Cómo estimaría las varianzas de dichos estimadores? [4]

---

[1] Una alternativa conveniente para aquellos que necesitan apoyo ver por ejemplo el Package "car" de R que permite responder a los requerimientos.

[2] Como ayuda, ver apartado 3.2, pág 53 del libro "*Practical Tools for Designing and Weighting Survey Samples*" (incluido en los textos del curso) y los comandos gamFit y gamEst, pág. 29 del manual del package "Practools" de R para determinar la estimación de los parámetros.

[3] Emplear el Package "Sampling" de R y los comandos "strata" y "getdata" para crear el archivo de la muestra.

[4] El apartado 2 del documento que acompaña al TP 1, "Qualite-Tille (2008).pdf", brinda un tratamiento inicial para estimadores de cambios en encuestas repetidas y de la estimación de sus varianzas que pueden ser de utilidad para dar respuesta al punto 6.