Post-Stratification

April 20, 2022

1 Notación

Para ser consistentes a lo largo del documento, vamos a utilizar la siguiente notación para todos los métodos que vamos a describir:

$\{1, ..., k, ..., L\}$: Conjunto de subpoblaciones.

$\{1, ..., i, ..., n\}$: Conjunto de personas entrevistadas.

d_i: tamaño de la red de contactos de la persona i.

y_{ik}: número de personas que conoce la persona i de la subpoblación k.

N_k: tamaño de la subpoblación k.

N: tamaño de la población total.

n_k: tamaño de la subpoblación k en la muestra.

N_h: el tamaño de la población oculta que queremos estimar. Usaremos la letra h para todo lo referente a la población oculta (*Hidden Population*).

2 Post-Stratification y su aplicación

La post estratificación es un método que funciona muy bien con encuestas multipropósito, de cara a poder tener un gran número de variables sobre las que realizar la post estratificación, pudiendo obtener así unas divisiones lo más buenas posibles del espacio para maximizar la precisión de nuestros resultados. En nuestro caso particular sólo tenemos datos de la región, por lo que es la única post estratificación posible. Las fórmulas a utilizar son:

$$(\bar{Y}_{ps})_{est} = \sum (N_k/N)\bar{y}_k,$$

donde $\bar{y}_k = \sum_{i \in s} Y_{ki}/N_k$ e $\bar{Y}_k = \sum_i Y_{ki}/N_k$. Además Y_{ki} corresponde con la estimación de N_h que realiza el individuo i de la subpoblación k.

Vemos que el fin de este método, es asignarle un peso de $1/N$ a cada uno de los individuos de la población, luego en función de las variables elegidas para

los estratos esto quedará mejor o peor representado. Una prueba rápida de que esto es así se obtiene al sustituir $\bar{Y}_k$ en $(\bar{Y}_{ps})_{est}$.

$$(\bar{Y}_{ps})_{est} = \sum_{k=1}^{L}(N_k/N)\bar{Y}_k = \sum_{k=1}^{L}(N_k/N)\sum_i Y_{ki}/N_k = \frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k} Y_{ki}.$$

Luego si encuestaramos a toda la población sería lo mismo que asignarle un peso de $1/N$ a cada individuo, como ya indicamos. Sin embargo, esto no suele ser posible ya que nuestro estudio se realiza sobre una submuestra de la población, por lo que usamos la media de las estimaciones en cada estrato como la estimación del valor verdadero de ese estrato, y le asignamos el peso correspondiente en función de la proporción de población que represente dentro de la población general. Para utilizar este tipo de métodos para los nuevos estimadores que conocemos, en vez de usar la media descrita por $\bar{y}_k$ usamos la estimación conjunta de N_h en dicho estrato, variando según el método, ya que usar la media sería el PIMLE o MoS, mientras que en el MLE serían el ya conocido cociente de sumatorios.

2.1 Estimadores y cómo implementarlo

2.1.1 Estimador $PIMLE$ (Plug-in Maximum Likelihood Estimator)

Este estimador surge de maximizar la función de verosimilitud asociada al método NSUM respecto a d_i, lo que nos lleva a la siguiente fórmula:

$$\hat{d}_i = N \cdot \frac{\sum_{k=1}^{L} y_{ik}}{\sum_{k=1}^{L} N_k}. \tag{1}$$

Esta formula nos permite la estimación del tamaño de la red de cada una de las personas que hemos encuestado, de forma que haciendo un *plug-in* en ele stimador básico obtenemos el estimador:

$$\hat{N}_h^{PIMLE} = \frac{1}{n}\sum_{i=1}^{n} N \cdot \frac{y_{ih}}{\hat{d}_i}. \tag{2}$$

2.1.2 Estimador MLE (Maximum Likelihood Estimator)

Killworth et al.[Killworth et al. (b)] propone una modificación de 2 de forma que, en vez de maximizar la función de verosimilitud respecto a d_i, lo hacemos respecto a N_h dónde los d_i están fijos, utilizando los estimados mediante 1 ($\hat{d}_i$). De esta forma, obtenemos el siguiente estimador:

$$\hat{N}_h^{MLE} = N \cdot \frac{\sum_{i=1}^{n} y_{ih}}{\sum_{i=1}^{n} \hat{d}_i} = \sum_{i=1}^{n} y_{ih} \frac{\sum_{k=1}^{L} N_k}{\sum_{i=1}^{n}\sum_{k=1}^{L} y_{ik}} \tag{3}$$

La gran diferencia en la estimación de la población desconocida que encontramos entre el $PIMLE$ y MLE es que el primero hace la media entre los diferentes encuestados mientras que el segundo maximiza la verosimilitud utilizando todos los datos de los que han respondido a la encuesta simultáneamente.

2.1.3 Estimador MoS (Mean of Squares)

Se parece mucho al que se muestra en el PIMLE pero utilizando un método diferente para calcular los d_i, ya que en vez de contarlos todos a la vez, vamos a hacer la media de todos ellos:

$$\hat{d}_i = \frac{\sum_{k=1}^{L} \frac{y_{ik}}{N_k}}{L} N. \tag{4}$$

De esta forma, si una población es muy grande, o en su defecto muy pequeña, podremos ver el impacto de la misma por igual junto a las demás, en vez de premiar que las poblaciones de estudio sean grandes. Este método nos permite tener en cuenta todas las subpoblaciones por igual, por lo que a pesar de no estar basado en un método de máxima verosimilitud, aporta una visión que es muy interesante y que funciona bastante bien. Para estimar N_h usamos la siguiente fórmula:

$$\hat{N}_h^{MoS} = \frac{\sum_{i=1}^{n} \frac{y_{ih}}{\hat{d}_i}}{n} N \tag{5}$$

2.1.4 Estimador $GNSUM$ (Generalized Scale-Up Stimator)

FALTA

2.1.5 Estimadores con pesos MoS, MLE y $PIMLE$

Podemos implementar a estos métodos los pesos de post-estratificación diferentes formas. Esta es la propuesta por el artículo [Habecker et al.]. En mi opinión el peso para el MLE también debería afectar al tamaño de la red individual d_i, es decir, al denominador. Usando los pesos del artículo [D. Holt et al.] ($w_i = N_k/N$, donde el individuo i pertenece a la subpoblación k) nos queda:

$$\hat{N}_h^{PIMLE} = \frac{1}{n} \sum_{i=1}^{n} N \cdot \frac{w_i \cdot y_{ih}}{\hat{d}_i} = \frac{1}{n} \sum_{k=1}^{L} N_k \sum_{i \in s} \frac{y_{ih}}{\hat{d}_i} \tag{6}$$

$$\hat{N}_h^{MLE} = N \cdot \frac{\sum_{i=1}^{n} w_i \cdot y_{ih}}{\sum_{i=1}^{n} \hat{d}_i} = \frac{\sum_{k=1}^{L} N_k \sum_{i \in s} y_{ih}}{\sum_{k=1}^{L} \sum_{i \in s} \hat{d}_i} \tag{7}$$

$$\hat{N}_h^{MoS} = \frac{\sum_{i=1}^{n} \frac{w_i \cdot y_{ih}}{\hat{d}_i}}{n} N = \frac{1}{n} \sum_{k=1}^{L} N_k \sum_{i \in s} \frac{y_{ih}}{\hat{d}_i} \tag{8}$$

3 Problemática y soluciones

El problema que tenemos al fijarnos en las preguntas de la encuesta es que no se realizan en el contexto de España, si no que se realizan en un contexto regional/autonómico. De esta forma, los estimadores que tenemos van a realizar una estimación precisa de la región en la que se encuentran, no del agregado

español. Como consecuencia no podremos utilizar estas preguntas directamente para utilizar el método descrito a lo largo de este documento. Soluciones:

- **Sin Post-Stratification**: Se podría hacer mediante la media ponderada de dos estimaciones diferentes. La primera utilizando los datos nacionales, obteniendo así una primera aproximación, $\widehat{N}_h$, mediante el tipo de estimador que queramos. Podemos realizar una segunda aproximación tomando la suma de cada una de las provincias/comunidades para las que tenemos datos, de forma que hacemos una estimación para cada una de las regiones y las sumamos para obtener el resultado total. Usaremos en nuestro estimador N_k en vez de N que "sería lo mismo" que aplicarles el peso N_k/N a la estimación global con estos mismos datos. Veamos un ejemplo con el MoS:, donde el estimador general para la población N sería:

$$\hat{N}_h^{MoS} = \frac{\sum_{i=1}^{n} \frac{y_{ih}}{\hat{d}_i}}{n} N$$

Mientras que un estimador para la población local sería:

$$\hat{N}_{kh}^{MoS} = \frac{\sum_{i=1}^{n} \frac{y_{ih}}{\hat{d}_i}}{n} N_k$$

Vemos pues que el paso de una a otra es la aplicación del factor descrito:

$$\frac{N_k}{N} \hat{N}_h^{MoS} = \hat{N}_{hj}^{MoS}$$

Sin embargo no podemos hacer esta relación para pasar de uno a otro, ya que cada uno de ellos mide una cosa diferente, pasar del general al local o del local al general sería pensar que todas las regiones de España tienen el mismo índice de Covid, cuando esto depende mucho de la provicia/comunidad. Luego las variables y_{ih} y d_i tienen que adaptarse a estas condiciones de ser nacionales o provinciales. Como tenemos condiciones provinciales, lo que tenemos que hacer pues es tener en cuenta esto y hacer un agregado que sea consecuente a nuestra situación. Tenemos pues que la opción más coherente sería estimar el tamaño de la población con Covid-19 en cada provincia y sumar todas las estimaciones para obtener la estimación a nivel nacional.

$$\sum_{k=1}^{52} N_{kh} = N_h$$

El estimador final sería una media ponderada de los dos anteriores. Sin embargo esto no tendría nada que ver con el proceso de post estratificación y creo que esta fue la solución aportada por Antonio. El agregado final quedaría de la forma:

$$\widehat{\widehat{N}}_h = w_1 \widehat{N}_h + w_2 \sum_{j=1}^{52} \widehat{N}_{kh}$$

- **Con Post-Stratificacion:** La solución alternativa que proponemos ([D. Holt et al.]) pasa por tener datos nacionales, es decir, habría que cambiar las preguntas:
  - *¿A cuántas personas conoce personalmente en esta región?*
  - *Que usted sepa, ¿cuántas de las personas que conoce en esta región han sido diagnosticadas o han tenido síntomas compatibles con COVID-19?)*

  Por las siguientes preguntas:
  - *¿A cuántas personas conoce personalmente en España?*
  - *Que usted sepa, ¿cuántas de las personas que conoce en España han sido diagnosticadas o han tenido síntomas compatibles con COVID-19?*

  De esta forma, si que podremos utilizar nuestro método de post estratificación, ya que todos los datos que obtenemos (y_{ik}, y_{ih}, d_i) son del contexto nacional, por lo que tiene sentido hacer estratos aplicando los pesos ya descritos. De esta forma nos quedaría el mismo estimador que en el apartado anterior, de ahí la discusión de la reunión, solo que cada uno de ellos expresan dos magnitudes diferentes.

$$\sum_{k=1}^{52} \widehat{N}_{kh}^{ps} = \widehat{N}_{h}^{ps}$$

  Como ya es tarde para realizar este paso, deberíamos encontrar una forma de extender la estimación de la población regional a nacional, lo cúal creo que es una tarea que no va a tener mucho éxito a nivel de mejora de la estimación, ya que sólo tenemos una decisión posible de post estratificación: la geográfica. Esto limita mucho el método y seguramente nos lleve a un trabajo en vano.

  Otra motivación de cambiar esas preguntas sería la introducción de sesgos más acentuados, en el sentido de que una mala división de la población en el primer método nos puede llevar a cometer errores grandes en la estimación. Por ejemplo, en este caso mucha gente puede conocer gente con covid de otras regiones pero no puede declararlos. En este caso la red social no parece estar determinada por el factor geográfico, pero no conviene usar el primer método de cara a no hacer divisiones demasiado pequeñas que condicionen totalmente la encuesta.

4 Conclusiones

La única diferencia entre estos dos métodos expuestos es la base que tienen y sus objetivos. Mientras que uno estima la población general mediante la suma de

las estimaciones de las diferesntes zonas del país, el otro usa los datos nacionales estratificados por región para hacer esta estimación lo mejor posible.

La principal diferencia es que el método con post estratificación nos abre la puerta a implementar numerosas mejoras (aunque es necesario un estudio más profundo de ciertas características de la población) ya que su principal motivación es tener diferentes formas de realizar la partición en estratos de la población nacional. Sin embargo, teniendo los datos sólo de las provincias/C.C.A.A no creo que la post estratifición aporte unos resultados esclarecedores o mucho mejores que los aportados por el método sin la misma, ya que el punto fuerte del método expuesto es tener varias poblaciones y la decidir cúales utilizar con el fin de optimizar la precisión de los resultados, pero en este caso estamos limitados por la falta de datos.

Para ver la utilidad real de la post estratificación de cara a siguientes estudios estaría muy bien implementar estructuras coherentes en las simulaciones y hacer todo este procedimiento, comparando los casos.

References

[Ian Laga et al.] Ian Laga, Le Bao, Xiaoyue Niu. (2020/21) Thirty Years of The Network Scale up Method

[Appendix] Ian Laga, Le Bao, Xiaoyue Niu. (2020/21) Thirty Years of The Network Scale up Method Appendix

[Habecker et al.] Habecker, P., Dombrowski, K., and Khan, B. (2015). Improving the network scale-up esti- mator: Incorporating means of sums, recursive back estimation, and sampling weights. PloS one, 10(12).

[Killworth et al. (a)] Killworth, P. D., Johnsen, E. C., McCarty, C., Shelley, G. A., and Bernard, H. R. (1998a). A social network approach to estimating seroprevalence in the united states. Social networks, 20(1):23–50.

[Killworth et al. (b)] Killworth, P. D., McCarty, C., Bernard, H. R., Shelley, G. A., and Johnsen, E. C. (1998b).

Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. Evaluation review, 22(2):289–308

[D. Holt et al.] D. Holt and T. M. F. Smith (1979). Journal of the Royal Statistical Society. Series A (General), Vol. 142, No. 1,pp. 33-46

[Ehsan Zamanzadea et al.] Ehsan Zamanzadea, Xinlei Wangb (). Estimation of population proportion for judgment post-stratification

[Feehan and Salganik] Feehan, D.M. and Salganik, M. J. (2016). Generalizing the network scale-up method: a new estimator for the size of hidden populations. Sociological methodology, 46(1):153–186.