Feehan and Salganik (2014 and 2016)

Suggest the use of GNSUM estimation techniques where there is non-random social missing and imperfect awareness about membership in the hidden population, and where more complex sample designs and incomplete sampling frames are used for data collection (Feehan and Salganik, 2016a). Two methods are proposed by these authors, and it is the second – the application of interpretable adjustment factors to the basic scale-up estimator- that is applied in this paper.

This approach to estimation builds upon the core insight that 'ordinary people have embedded within their personal networks information that can be used to estimate the sizes of hidden populations' (Feehan and Salganik, 2016a p.154).

Method

The estimation method is built upon the foundations of aggregated relational data (ARD) which 'asks respondents how many individuals they know in a particular group of interest' (McCormick et al., 2012, p.179). Researchers view the number known in a group of interest as a proportion of the respondent's network (which requires estimating the respondent's total network size and then 'scale up' from the total proportion of respondents' networks to the size of the group of interest in the overall population. (op cit)

In our survey we collected *such aggregated relational data* (McCormick et al., 2012) about a group of known size by asking respondents the question 'do you know personally of any other domestic workers who are engaged in domestic work in the UK? And if so, how many? '.

As we believed that many of the domestic workers who would respond to our survey may have contacts with other domestic workers from their home countries who may have chosen to migrate for work elsewhere, our question was explicitly designed to eliminate those domestic workers known to the respondent who may have found work in other destination countries.

McCormick et al. (2012) describe three types of bias in mental link tracing estimates such as the Network Scale Up Estimation that we used as part of our estimation process: barrier effects (where some individuals know systematically more or fewer members of a specific subpopulation than would be expected under random mixing (e.g. as a result of know more people of a specific age or gender); calibration bias – when respondents have difficulty in recalling accurately the number of members of a group that they know and preferential nomination bias – more prevalent in an alternative form of research design where egocentric nominations are not made randomly –and so of less relevance to our study - which may result in a subset of alters which are not representative of the overall set of individuals that they know in that group.

Of relevance to our approach, we asked domestic workers to estimate the number of other domestic workers whom they knew within their networks by referring to the contact details held within their mobile phones. This, we believe, may have reduced the calibration error identified by McCormick et al., 2012) while also having the effect of

Discussion

The social network structure of our sample (Figure x) …

r

Research limitations

To provide size estimates from aggregate relationship data, researchers have used the basic scale up estimator – first proposed by Killworth et al. (1998b). This requires two main building blocks:

1) aggregate relational data about the hidden population – used to estimate the number of connections that respondents have to the hidden population

How many domestic workers do you know that experience this form of exploitation?

2) aggregate relational data about the groups of known size are used to estimate the number of connections that the respondents have in total.

How many domestic workers do you have listed on your mobile phone?

This estimate is made using the known population estimator (Killworth et al, 1998a)

Simple sum of number of know people with connections to those experiencing exploitation/ total number of domestic workers' known.

The problem may be that actually there are many overlapping communities, due to the horizontal nature of the data sets. The data sets are not discrete.

What is an alter sample? – goes beyond the respondents to consider the networks of those to whom they are linked.

" Researchers who desire absolute size estimates multiple the alter sample proportion to the size of the entire population" p.2

E.g. In the year to December 2022, there were 18,553 overseas domestic workers visas granted to migrant domestic workers to work in the UK

[If the chances of a connection being shared between two of your respondents is independent of the characterstic (the same chance of know) then the overall proportion estimate is the same whether or not there is overlap

The NGO introducing them increases the overlap, but doesn't increase the probability of a connection happening outside of the characteristic. Doesn't increase the chance of them working

The people that they meet through the NGO aren't representative of the people that they meet in work. This increases the apparent proportion of exploited people.

Basic scale up estimation is derived from the basic scale-up model, the assumptions of which have been proved to be problematic –

1) social ties are formed at random

2) respondents are perfectly aware of the characteristics of their alters

3) respondents are able to provide accurate answers to survey questions about their personal networks

In particular, violations of the random mixing assumption is referred to as barrier effects

Violations of perfect awareness are termed transmission error while respondent accuracy problems are termed recall error.

Advocate a different approach to improve scale up estimation – deriving the new, generalised network scale-up estimator from a simple identity.

Inspired by earlier research on multiplicity estimation and indirect sampling reveals that researchers can produce a size estimate by combining aggregate relationship data collected from the frame population with similar data collected from the hidden population

What aggregated relationship data do we need to collect from the frame population – and how do we do this?

The advantage of this is it is not biased by barrier effects, it naturally accounts for imperfect social awareness (transmission effects) it accommodates incomplete sampling frame and complex sampling designs

Second reading

Generalised scale up estimator

Requires two properties; a property of the frame population and a property of the hidden population

{the frame population is the number of domestic workers in the UK whom it is possible for us to include in our sample)

[Trust for London estimate that there were 16,000 domestic workers in the UK in 2022,]. According to the ONS, this has risen to 18,553 by December 2022

A GNSUM estimate may be developed either from adding a small number of questions to already planned studies or through the application of 'interpretable adjustment factor's that may be applied to the basic scale-up estimator

In our study we have **aggregate relational data** from our frame respondents about two groups:

The number of domestic workers that each respondent knows

The number of these domestic workers experiencing (various forms of exploitation)

In estimation using basic scale-up data has two main building blocks: First aggregate relational data about the hidden population (i.e. those experiencing exploitation). Second, aggregate relational data about the groups of known size (i.e. those in domestic work) are used to estimate the number of connections that respondents have in total. This estimate is made using the known population estimator (Killworth et al, 1998a)

Total number of connections to exploited workers/ total number of connections in total.

Researchers who design absolute size estimates multiply the alter sample proportion by the size of the entire population – assumed from official statistics (i.e. 18,553 domestic workers)

However, the basic scale up estimation suffers from problematic assumptions described during my first reading, as barrier effects, transmission error and recall error.

Generalised Network Scale Up Method rests on the understanding that researchers can produce a size estimate by combining aggregate relational data collected from the frame population with similar data collected from the hidden population.

Improved estimation due to lack of bias from barrier effects and transmission error – recall errors still possible – and can accommodate incomplete sampling frames and complex sampling designs.

Does require data to be collected from the hidden population. **[Can I say that responses from those respondents experiencing abuse are members of the hidden population? – I don't think so. Instead, we need to complete some further data collection from those who are reporting exploitation about how many other domestic workers to whom they are connected know about the exploitation they have experienced. There are potential language barriers here – the question may need to be translated into the preferred language of the respondent]**

Section 2 – derives the gNSUM and describes the data collection procedures needed to use it.

Section 3 – examines the difference between Generalized and basic scale up estimators introducing **3 factors to make the basic scale up estimator consistent and essentially unbiased .[Question 1 could these be used to improve our basic estimator?]**

Section 4 – introduces a new variance procedure for both estimators

Section 5- makes practical recommendations for the design and analysis of future scale-up studies

Section 6 – summary and outline of next steps

Section 2 – Generalised scale-up estimator

Very useful graph.

Yn,H = number of domestic workers in exploitation know by yn [we have this information]

vnU = we don't have this information from the hidden population [i.e. all those in the frame population who report exploitation – we need to ask them, how many of your domestic worker contacts know that you are exploited?]

Real studies sample from a subset of U called the frame population (e.g. adults)

Two types of reporting errors are possible: false positives and false negative. False negatives do not bias the estimation and false positives will introduction a positive bias into estimates, though these are believed to be rare [though this may not be the case in relation to exploitation] Appendix A shows how to adjust size estimates if the rate of false positives an be estimated – or, if this is not possible equation **Appendix A equation A7 can be used as a robustness check**

Collecting empirical data on outreports from the frame population and the average number of in reports from the hidden population.

yF,H = total number of out-reports can be estimated from aggregate relational data about connections to the hidden population.

We have this data – sum numbers of domestic workers in the frame sample reported to be in know situations of exploitation. Where Piei are the 'probabilities of inclusion from our sampling design' p.6) Review Horvitz-Thompson estimator (Sarndal et al., 1992) and Appendix B (Result B1)

Mean of v H,F can be estimated by asking members of the hidden population How many domestic workers do you know ? [which we have] and then how many of these domestic workers know that you are exploited? Which we need to collect

The probe alters group A is the concatenation of all these groups (we have just one group – DWs that you know – is this a problem?

[I am assuming that, if we know the total number of outreports – and can gather data about the average number of inreports from the hidden population – then we can produce an estimate of the number in the hidden population.

e.g. say each of our respondents identifies 2 dws that they know are experiencing exploitation and we have 97 respondents

total number of outreports = 2*97 = 194

And if the average number of connections is 1.5

Then the size of the hidden population is 194.]

Section 3 – relationship to the basic scale-up estimator

Basic scale-up indicator is (Equation 13) Which we have the data to calculate

Estimate of the hidden population = Sum of all items in the sample (yi,H) number of out reports of links to exploited domestic workers from person i

Divided by sum of all items in the sample estimate of di,u – the number of undirected network connections she has to everyone in U – (i.e. the number of domestic workers she knows)

Multiplied by N – where N is the population size (?)

[so, for example, sum of 2x97 /3 x 97 =6,272]

Feehan and Salganik suggest 3 adjustment factors which, if correctly applied, can convert a Basic network scale up measure into a measure of generalisable network scale up :

Frame ratio, degree ratio and true positive rate.

Where Nh = Basic scale up x 1/ frame ratio x 1/degree ratio x1/ true positive rate

Frame ratio (equation 17) is equal to average number of connection from a member of F to the rest of F (which we have)/ average number of connections from a member of U to F (which we could perhaps estimate?)

Degree ratio (equation 18) is equal to the average number of connections from a member of H to F (which we have) / average number of connections from a member of F to the rest of F (which we can estimate through referral phone number linkages)

True positive rate is equal to number of in reports to H from F (which we have)/ number of edges connecting H and F (could this be number of referral phone numbers to exploited people)? Maybe

Section 5 – practical recommendations

If you have samples from H and F, use

If just an F sample,

Suggest 4 recommendations:

Consider using a basic scale up estimator that removes the need to adjust for the frame ratio (this is useful, since this was the estimator that we needed to estimate). Instead, they propose the use of the basic scale up estimator in **equation 23**.

Estimate of the hidden population = estimate of yF,H / estimate of d F,F x Nf (see earlier approach under section 3 – which is in fact closer to this estimation) where Nf = sample population

Estimate of y F, H

Sum of all items in Frame sample : number of out reports of links to exploited domestic workers from person i…n

Divided by

Divided by sum of all items in the sample estimate of di,u – the number of undirected network connections she has to everyone in F – (i.e. the number of telephone connections listed between domestic workers)

Multiplied by NF – sample framework number

Where d F,F equals the total number of connections between adults and adults rather than estimate of d f, U the total number of connections between adults and everyone). In order to do so, researchers should design the probe alters for the frame population, AF so that they have similar personal networks to the frame population.

Second, be explicit about the values assumed for dF (degree ratio) and TF (True positive rate), equations for which are specified in Section 3.

Third, recommend that researchers assess the robustness of their estimates to any assumptions e.g. the implications of false positives. Further details are given in Appendix E

Fourth, provide confidence intervals using the rescaled bootstrap procedure given in section 4, described below:

Section 4: Variance estimation

Problems with the current variance estimation proposed by Killworth et al. (1998b) derived from the basic scale-up model result in too small an interval due to assumptions related to simple random sampling and the lack of recognition of more complex sampling designs.

Variance estimation – with a sample from F

1) generate B replicant samples by randomly sampling with replacement from sF

2) use these replicate samples to produce a set of replicate estimates Estimate NH1 … Estimate NHB

3) combine to produce a confidence interval, for example by the percentile method which chooses the 2.5^{th} and 97.5^{th} percentiles of the B estimates (Fig F.1) Efron and Tibshirani, 1993)

Perhaps take the same number B as of Primary Sampling units (seed responses) i.e selecting random samples with the number of primary responses – do this B times – say 10? Identify 2.5 and 97.5 confidence limits

If data from the hidden population is obtained: Instead, suggest that the best available bootstrap method for RD sampling data is that introduced in Salganik 2006. This requires researchers to divide the sample of the hidden population into two mutually exclusive groups. Feehan and Salganik recommend diving the hidden population in to those who are above and below the median of their estimated visibility vi,F in order to capture some of the extra uncertainty if there are strong tendencies for more hidden members of the hidden population to recruit each other.(p. 39

Notes: Appendix C1.

Need to use weighted sample means to estimate averages for the hidden population.

Initial Action plan

1. Using current data, conduct basic scale-up estimation and explain limiting factors.

2. Estimate frame, degree and true positive rates and make these adjustments to basic rate estimation to produce an approximation to GNSUM

3. Attempt to collect additional data from members of hidden population upon return from China (confirm date) to provide an alternative means of GNSUM estimation (?)

Send protocols for achieving this with Selim today – for discussion tomorrow

Prepare and submit paper for academic review prior to publishing in first or second-stage report.

Delay publication of GNSUM in first-stage report. First stage report to descriptive statistics of the frame population, second-stage paper to explore opportunities for further data collection from the hidden population.

Queries raise by Scott

What is the sample frame F? Is it the set of those working in DW in the UK

What is the hidden population, H? Is this the set of those working in DW being exploitated?

What are the probe alters? The total number of DWs known to the respondents

Is there a separate sample from H (e.g. using RDS/ Snowball sampling) RDS is a non-probalistic method of obtaining a sample where there is no sampling frame

Do we have _enriched_aggregate relational data F&S2016 p.161

I don’t see how one could calculate equation 7 in F&S. What am I missing?

I can see how one could calculate eqn 23 but I do not seen how the ‘improved version’ in eqn 24 could be calculated.

Aggregate relational data – one common method for sub-population size estimation (McCormick et al., 2012)

The list of potential alters a respondent could nominate is limited to people with whom the respondent has ties.

We have two types of aggregated relational data questions:

How many domestic workers do you know?

The true size of this population is unknown, but can itself be estimated using the number of overseas domestic worker visas issued in a year/2 [each visa is valid for 6 months]

How many domestic workers do you know who experience exploitation [in total, and of specific types]

ARD

Does not examine any of a respondent’s links directly. Advantages in that it ‘reaches’ more respondents and easily asks about multiple groups, but is limited by the need to estimate degree and may be susceptible to recall issues (calibration bias).

How many xs do you know?

How many domestic workers do you know?

How many domestic works do you know who experience x form of exploitation?

‘Know’ was more rigorously defined to include only those whom the respondent had listed on their mobile phone. This provides a summarised census of the entire respondents’ network

Research limitations

Enriched aggregate relational data (Feehan and Salganik, 2016a) was not collected from members of the hidden population. Feehan and Salganik define enriched relational data as ‘

A series of questions asked of the hidden population about their connections to certain groups and their visibility to these groups which may be elicited through techniques such as the game of contacts

References

McCormick, T., He, R., Kolaczyk, E., and Zheng, T. (2012). Surveying hard-to-reach groups through sampled respondents in a social network. Statistics in Biosciences, pages 1{19.

Feehan, D. M., & Salganik, M. J. (2016a). Generalizing the Network Scale-up Method: A New Estimator for the Size of Hidden Populations. *Sociological Methodology*, *46*(1), 153–186. https://doi.org/10.1177/0081175016665425