A new data collection strategy based on respondent groups and conditional response rates

A case from the Norwegian Labour Force Survey

Bengt Oscar Lagerstrøm[1], *Magnar Lillegård*[1] *and Tora Löfgren*[1]

1 Introduction

The Norwegian Labour Force Survey (LFS) is run as a panel survey in eight waves. The sample consists of about 12,000 family units (24,000 persons) each wave. All family members aged 15–74 participate in the survey, answering questions about their situation during a specified reference week. As from 1996 each family participates in the survey eight times during a two years period, i.e. one wave every three months.

The development in LFS recent years has been towards lower participation, and it has become increasingly difficult to get hold of the respondents. As survey response rates are decreasing both national and worldwide (e.g. de Leeuw and de Heer 2002) more resources are being used to increase survey response rate and/or dealing with the consequences of non-response (e.g. Bethlehem et al. 2011). Statistics Norway is in the bootblock for developing a more dynamic data collection system in order to avoid costs to completely derail based on available paradata to establish a responsive data collection design as described by Groves and Heeringa (2006). Based on the Norwegian LFS, Jentoft and Löfgren (2015) have shown pretty stable response propensities during the data collection. In this paper we want to look further on different respondent groups and how to arrange effective treatments based on knowledge of conditional response rates.

The analyses are based on two different data sets. We first look at response and non-response groups for persons with all eight waves inside the seven years period 2006–2012. By excluding those with missing value on one response variable or more, we end up with a data set of 60,608 persons. Then we compare this with data collected more recently, the period 2013–2015.

2 Measures of response and non-response

2.1 Conditional response rates

If you come into the LFS, you are most like to stay there. The 2006–2012 data show that of those who come into the survey in the first wave, 94 per cent will also take part in the second wave (Figure 1). If you come into the survey in the second wave, there is an 82.5 per cent chance that you will also take part in the third wave, and so on.

On the opposite side, if you are not in the survey in one wave, it is less likely that you will be there the following wave. For example, if you are not in the survey in the first wave, there is only 46 per cent chance that you will be there the second wave. And if you're not in the survey in the first *and* second wave, the chance is only 26.6 per cent that you will be there the third wave.

Data for the most recent period, 2013–2015, shows the same pattern but numbers are at a lower level, especially for the first and the second wave, see Figure 2. The third wave shows numbers more in line

[1] Statistics Norway, Box 8131 Dep., 0033 Oslo, Norway

with the bigger data set over a longer period. Comparing the two figures, we clearly see the fall in response rates.

Figure 1. Mean conditional response rates at the first, second and third wave. LFS 2006–2012. Blue colour means respondent, red colour non-respondent

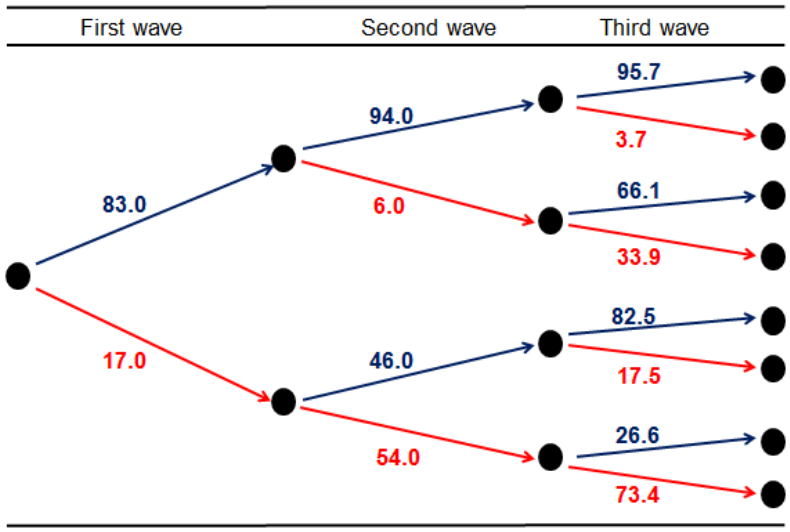
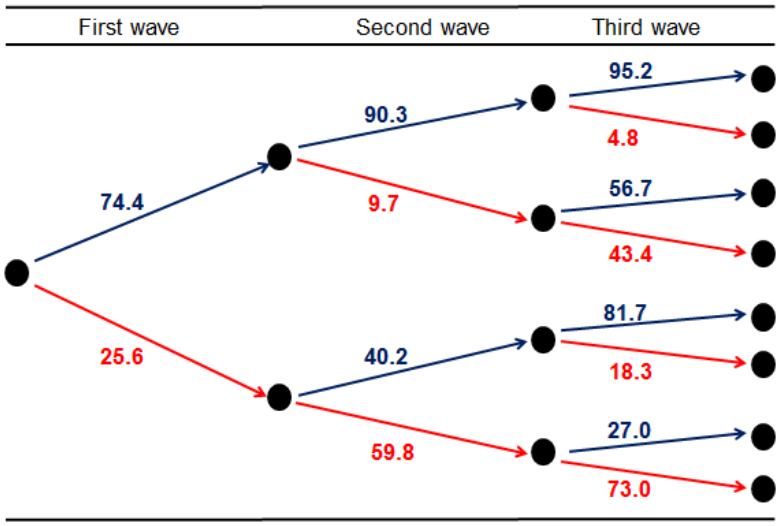


Figure 2. Mean conditional response rates at the first, second and third wave. LFS 2nd quarter 2013–2nd quarter 2015. Blue colour means respondent, red colour non-respondent



2.2 Respondent groups

For research purposes, we divide the respondents into five respondent groups.

1. Respondents: Take part in all eight waves
2. Late entries: Sample members participate in second wave or later, and then stay there
3. Wave non-response: Sample members do not participate in a particular wave of the study
4. Attritioners: Successive non-respondents of eligible respondents after start of panel ($t = 2, 3, ...$)
5. Non-respondents: Non-response of eligible respondents from $t = 1$ and for rest of the life span of the panel.

Figure 3 and Figure 4 below show the distribution of the respondent groups. By comparing, we see that in the latter the respondent group is smaller, while the late entry group is larger. This reflects the decreased response rates in first and second wave (Figure 1 and Figure 2). That means there is a tendency that people enter the survey at a later stage. We also notice that the wave non-response group is rather large. Both late entries and wave non-response have been shown by Stensson and Lagerstrøm (2010) to have a tendency to be self-reinforcing. A trend in this direction should therefore be handled carefully.

Figure 3. Distribution of respondent groups. LFS 2006–2012

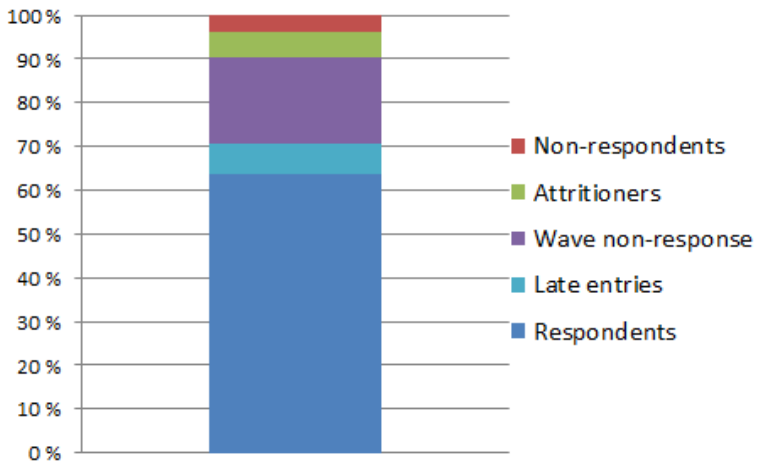
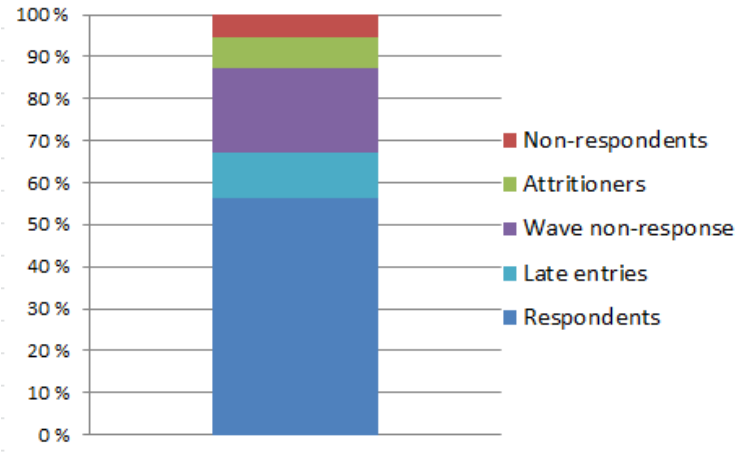


Figure 4. Distribution of respondent groups. LFS 2nd quarter 2013–2nd quarter 2015



The distributions shown are the estimated probabilities to be in a certain respondent group when we stand at the left node in the tree diagrams in Figure 1 and Figure 2. In principle, we can compute such distributions for every node. For example, when we stand at the bottom node in the second column (i.e. after the first wave), the probability of being in the respondent group is zero, and therefore the probability of being in one of the other four respondent groups must have increased. A model for the respondent group distribution can possibly be extended by taking demographic variables (gender, age, …) into account as well. An open question is whether we can use such information in our data collection strategy.

How respondent groups are correlated with age, gender, native born and employment is shown in Table 1. Typically we see that the respondents differ from the other respondent groups, i.e. higher age, lower proportion of men, higher degree native born and employed. With respect to the variable of

interest, employment, the attrition and the non-response group differ mostly from the respondent group.

Table 1. Statistics on age, gender, native born and employment, by respondent group. LFS 2006–2012

Respondent group	Mean age	Proportion of men	Proportion native born	Proportion register employed[2]	Proportion LFS employed[2]
Respondents	45.0	50.1	96.3	67.8	73.2
Late entries	41.9	52.5	91.3	63.3	69.0
Wave non-response	40.0	50.8	91.4	65.5	71.3
Attritioners	41.6	52.5	90.3	59.4	64.7
Non-respondents	41.4	57.5	84.8	56.9	-

3 New data collection strategies

3.1 Better contact information, fewer call attempts

To have interviewers searching for contact information costs a lot. Another treatment in terms of trying to get control of the money flow in the data collection has been our recently established access to a register with proper contact information for all citizens in Norway. The register retrieves contact information from citizen's digital contact with public government, including tax payers' declaration. This means that at least once a year the telephone numbers and email addresses are updated by the respondent and have the potential to be reliable and of good quality. The new register has been available during the spring. Piloting has shown promising results on contact rate, representativity and cost, but still we have to implement automatic tracing procedures for the LFS and other surveys to make sure that we use the potential in the register (Lagerstrøm and Wangen 2015). So fare, we have also learned that higher contact rates don't automatically leads to higher response rates, although several respondents are available for persuasion. One possible alternative is to use the time and money we have saved on call attempts on educating the interviewers even more in refusal conversation, or developing different information strategies for different respondent groups.

3.2 Special trained interviewers for all 'first-time respondents'

As demonstrated in Section 2.1, it is important that we get cooperation with respondents as early as possible in panel surveys. In the first wave Statistics Norway always uses special trained interviewers who are extra good in presenting the survey and making it attractive for the respondents to participate. They are also experts on refusal conversation. This means that attention is much greater on the respondents in the first wave than in subsequent waves. Because of the special trained interviewers used in the first wave, there is probably not much more to gain at that stage of the survey. So an extra effort should be made in the second wave. As shown, those who answer in the first wave are most likely to continue; therefore one should concentrate on the first wave non-responders. Among these, less than 50 per cent join the survey in the second wave. If that percentage could be increased, we would probably see a huge effect on the participation in the rest of survey.

So, one treatment we have come up with is to manipulate the usual data collection strategy by giving those who have not responded in the first wave the same attention (or 'treatment') as initial first-wave

[2] Per cent employed of population 15–74 years old, *not* per cent of labour force.

respondents; i.e. the special trained interviewers. Our hypothesis is that we can change the respondent's behaviour and affect level end of attrition, wave non-response, late entries and non-response, with early efforts. This experiment has just started, so it is too early to see any effect of it in data.

3.3 Responsive design

We wish to implement responsive design in the data collection at Statistics Norway, but among other things we need more treatments in place. In this paper we have defined five respondent groups, and we raise the question whether modelling of respondent groups can be useful in our data collection. One task that lies ahead is inventing different treatments and then test out which treatment has an effect on what group, in order to avoid the non-response, attrition or that people gradually participate later and later in the survey.

At the same time, it is important to have a web version of the questionnaire ready and set into production. With today's technology and smart phones, it is not unthinkable that some of those not participating in the study today would participate if they could report on their own and not having to talk to an interviewer on phone.

4 Some topics to discuss

- How can we prevent that a person ends up in the late entry group instead of the respondent group?
- Is it to possible to differentiate the data collection strategies based on the identification of respondent groups and conditional response rates?
- What other possible treatments could be included in a responsive design (apart from different modes)?

References

Bethlehem, J., F. Cobben, and B. Schouten (2011): *Handbook of nonresponse in household surveys*, New York: Wiley.

De Leeuw, E. and W. de Heer (2002): 'Trends in household survey nonresponse: a longitudinal and international comparison', in R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little (eds.): *Survey Nonresponse*, New York: Wiley, 41–54.

Groves, R. M. and S. G. Heeringa (2006): Responsive design for household surveys: tools for actively controlling survey errors and costs, *Journal of the Royal Statistical Society A* **169**, 439–457.

Jentoft, S. and T. Löfgren (2015): Response propensities in the Norwegian Labour Force Survey: The move towards a responsive design (unpublished).

Lagerstrøm, B. O. and G.-E. Wangen (2015): Erfaringer med Kontakt- og reservasjonsregisteret i Statistisk sentralbyrå (unpublished)

Stensson, S. and B. O. Lagerstrøm (2010): The effects of longitudial 'drop-outs' and late respondents on the main EU-SILC indicators. Paper presented at the Nordic Statistcal Meeting, Copenhagen 2010.