Combining information from different survey samples - a case study with data collected by world wide web and telephone

Magne Aldrin
Norwegian Computing Center
P.O. Box 114 Blindern
N-0314 Oslo
Norway
E-mail: magne.aldrin@nr.no

Summary

This paper considers an investigation of the users’ opinion of the web site ODIN, which is the official web site for information from the Norwegian Government and Ministries. A survey were performed by asking people who log on to the ODIN web site to fill out questionnaires on their use of the web site. This could be a potentially very useful way to collect information about the users’ needs, since the sample was selected directly from that subpopulation who uses the web site. However, answering was voluntary, and most users did not answer the questionnaire. The large proportion of possibly informative non-response makes it difficult to draw general conclusions from the observed web data only. Therefore, a telephone survey was carried out in addition, with people randomly selected from the whole population in Norway. The information from the web survey and the telephone survey were combined to give reliable results. A crucial part of the statistical modelling was to take into account the different sampling schemes of the two surveys: In the web survey, people were sampled proportionally to their frequency of using ODIN, whereas people in the telephone survey were sampled with equal probability. The statistical approach was Bayesian, and the BUGS software was used to perform inference.

Keywords: unequal sampling probabilities, Markov Chain Monte Carlo, BUGS

1 Introduction

Sample surveys have traditionally been performed by personal interviews or by a questionnaire on paper. Often the sample selection are controlled by the investigator, for instance by a complete randomized sampling from a population, or perhaps by stratified sampling. The Internet revolution has given opportunities to other, and cheaper, ways of collecting survey data. Personal questionnaires send by e-mail have been used for a while (Bachmann, Elfrik and Vazzana 1996). Recently, surveys have been performed by asking people who log on to certain (world wide) web sites to fill out questionnaires about their use of the web site. This is a potentially very useful way to collect information about the users needs, since the sample is selected directly from that subpopulation who uses the web site. However, if answering is voluntary, one could expect that most users neglect to answer the questionnaire. The large proportion of possibly informative non-response would make it difficult to draw general conclusions from the observed data only.

In 1997, the Norwegian Computing Center has performed an investigation on a web site called ODIN, which is the official web site for information from the Norwegian Government and Ministries (Solheim and Tjøstheim 1997). In a period of 35 hours, each 20-th ODIN user was asked to fill out a questionnaire. There were 1130 possible responders, but only 131 (11.6%) of these answered, which certainly is very little for making inference about the entire population of users. The sample may potentially be very biased, since we do not know anything about who answers and who does not. In order to try to correct such information, a telephone survey was carried out in addition. 1012 persons, randomly selected from the whole population (at least 15 years old) in Norway, were rang up. Of these, 34 (3.4%) had used ODIN more than once.

In this paper, I combine information from the web and telephone surveys. I will focus on estimating the proportion of frequent ODIN users and of satisfied ODIN users among the subpopulation consisting of those people in Norway who have used ODIN more than once. In addition, I am interested in *weighted* proportions, where each user is weighted proportionally to how often she uses ODIN. The rationale behind this is that when trying to improve ODIN, it may be more important to satisfy frequent users' wishes than those of sporadic users. On the other hand, less frequent users may become frequent users in the future if ODIN is improved according to their needs, therefore the *unweighted* proportions are also of interest.

The members of the subpopulation of interest (those who have used ODIN more than once) are classified according to their frequency of using ODIN and to their opinion regarding their benefit of using ODIN. The definitions are

- frequent user: uses ODIN “daily” or “weekly”
- less frequent user: uses ODIN “monthly”, “periodically” or “a few times”
- satisfied user: has found ODIN “very useful”
- less satisfied user: has found ODIN to be “useful” or of “little value” or “not useful”.

The users are then cross-classified into the following four categories:

category 1: less frequent and less satisfied users
category 2: frequent and less satisfied users
category 3: less frequent and less satisfied users
category 4: frequent and satisfied users.

In this paper, I concentrate on only these few possibilities, because my aim is to illustrate the methodological approach. Of course finer subdivisions, taking into accounts more questions etc., could be built.

We are interested in the proportion of people within each category, and especially some marginals: those of frequent users and of satisfied users. Consider all the users in the subpopulation (of more-than-once users). The *unweighted* proportion of satisfied users is defined by

$$\text{proportion of satisfied users} = \frac{\text{number of satisfied users}}{\text{total number of users in subpopulation}} . \qquad (\ 1)$$

Let now f_j be the j-th users frequency of using ODIN, that is number of accesses per time unit. The corresponding *weighted* proportion of satisfied users is defined by

$$\text{weighted proportion of satisfied users} = \left(\sum_{j \in \text{ satisfied users}} f_j \right) \Big/ \left(\sum_{j \in \text{ all users in subpop.}} f_j \right) . \qquad (\ 2)$$

The *unweighted* and *weighted* proportions of frequent users are defined similarly.

Since people in the telephone survey are sampled with equal probability, the *unweighted* proportion of satisfied users may be estimated from the telephone data set alone, using a sample version of (1) to the telephone data. However, by using additionally information from the web sample, the estimate of the *unweighted* proportion may possibly be improved via what we may call Bayesian rescaling. On the other hand, in the web survey, each person is sampled with a probability proportional to her frequency f_j. Therefore, if there were no non-response in this sample, one could get unbiased estimates of the *weighted* proportion of satisfied users (2) simply by using the sample version of (1) to the web data, without knowledge of the individual frequencies f_j. Unfortunately, since most of those that logged on did not answer, inference based only on the observed answers could be very misleading exactly because the probability for a user not to respond may depend on which of the four categories she belongs to. Hence, to get reliable results for the *weighted* proportions, it is essential to perform a combined analysis of the two data sets.

The statistical approach is Bayesian, and the BUGS software (Spiegelhalter et. al. 1996) is used to perform inference.

2 The two survey data sets

2.1 Telephone data set

The telephone data were collected by Norsk Gallup. The sampling has been done according to their ordinary routines. Usually, when analysing such samples, Norsk Gallup corrects for possible bias on certain criteria, such as sex, age and residence. However, for simplicity we will here assume here that these interviewed candidates are randomly chosen from the Norwegian population.

This data set consists of 1012 persons. They were initially asked if they had used Internet, and if they answered yes, they were asked about ODIN. Of these, 34 answered that they had used ODIN more than once, and furthermore answered to a question related to their benefit of ODIN. The 34 responses are classified into one of the categories 1 to 4. The remaining persons never use ODIN or did it only once. These are said to belong to category 0. Table 1 shows the absolute numbers within each category.

Table 1 Number of persons within each category in the telephone data set**.**

Category	0	1	2	3	4
Number	978	16	10	5	3

2.2 Web data set

The web data set was collected from ODIN users over a period of about 35 hour. In this period, every 20-th user who logged in was asked if she was willing to answer the questionnaire. If she answered no, she was connected directly to the ordinary ODIN pages. 1130 persons got the question. Of these 998 refused to answer the questionnaire, and one failed to fill out the questionnaire correctly. These 999 persons are considered as missing observations. Of the remaining 131 persons, 10 were using ODIN for the first time. These 10 are classified to category 0. Remember that in the telephone data set, category 0 also contained persons who had never used ODIN, but such persons were not sampled through the web survey. The remaining 121 persons were classified into one of the possible categories 1 to 4. Table 2 shows the absolute numbers within each category.

Table 2 Number of persons within each category in the web data set**.**

Category	0	1	2	3	4	missing
Number	10	39	20	32	30	999

3 Model specification

Consider first the people reached by the telephone survey. Let p_i, i=0, ... , 4 be the probability for such a person belonging to the i-th category, i.e.

$$p_i = P(\text{person reached by phone belongs to category i}) \qquad i = 0, \ldots, 4\,. \qquad (3)$$

Next, consider the web survey. Let q_i denote the corresponding probabilities

$$q_i = P(\text{person reached by web-questionnaire belongs to category i}) \qquad i = 0, \ldots, 4 \qquad (4)$$

The rationale behind the use of two different probabilities for the two data sets is that the sampling procedures differs among the two surveys. In both surveys, the persons are sampled from the total Norwegian population. In the telephone survey, each person is sampled with equal probability, while by the web survey each person is sampled proportionally to her frequency of using ODIN. If a person uses for instance ODIN daily, the probability of getting the web-questionnaire would be about 30 times higher than if she would access ODIN say on a monthly basis.

The next step in our modelling consists in introducing dependency between (p_0, ..., p_4), and (q_0, ..., q_4) which are of course related parameters. Consider the total population, and let n_i denote the number of persons in the i-th category. Let w_{ij}, j=1, ..., n_i be weights proportional to the frequency of using ODIN for the j-th person of category i. The sampling schemes then link the p- and q-probabilities together through

$$q_i = (1/n_i)\sum_{j=1}^{n_i} w_{ij}p_i = w_i p_i \qquad i = 0, \ldots, 4, \tag{5}$$

where w_i is proportional to the average frequency within category i. The w_i's are normalised such that $\Sigma q_i=1$. Of the total population, most persons in category 0 never use ODIN. These are sampled with probability zero in the web survey, so that they have zero weight and they do not contribute to the q-probabilities.

The definitions of p_i and q_i in (3) and (4) follows from the sampling schemes. However, they have also alternative interpretations:

- p_i is the proportion of the total population that belongs to category i, i=0, ..., 4.
- q_i is the corresponding *weighted* proportion, where each person is weighted proportionally to her frequency of using ODIN.

Members of category 1 and 3 are less frequent users, and it is reasonable to assume that the average use are approximately the same in the two categories. To be parsimonious we therefore assume $w_1=w_3$, and for the same reason we assume $w_2=w_4$. The averaged frequencies are ordered such that $w_0<w_1<w_2$, as a consequence of the division into the three levels of frequency. This gives inequalities between p and q such as $(p_2/p_1)<(q_2/q_1)$. Since both the p_i and q_i probabilities add to one, there are now 6 free parameters to be estimated.

In order to treat easily the ordering of the weights w_i, we reparameterize them using c_0, c_1 and c_2 defined as follows

$$w_0 = c_0 \qquad , \qquad w_1 = c_1c_0 \qquad , \qquad w_2 = c_2c_1c_0\,, \tag{6}$$

and assume that

$$c_0 > 0 \qquad , \qquad c_1 > 1 \qquad , \qquad c_2 > 1\,. \tag{7}$$

Now, p_i and q_i are linked together through

$$q_0 = c_0 p_0 \quad , \quad q_1 = c_1 c_0 p_1 \quad , \quad q_2 = c_2 c_1 c_0 p_2 \quad , \quad q_3 = c_1 c_0 p_3 \quad , \quad q_4 = c_2 c_1 c_0 p_4 \,. \qquad (8)$$

As mentioned, it is reasonable to assume that the probability for non-response in the web sample varies with the category. A model with one specific probability of non-response for each category would be difficult to identify. Instead, we model the probability of non-response by a logistic regression equation with four parameters. Let m_j be the probability that person j in the web sample does not answer. The model is

$$\log\left(\frac{m_j}{1 - m_j}\right) = \beta_R + \beta_F F_j + \beta_U U_j + \beta_N N_j \qquad j = 1, \ldots, 1130 \,. \qquad (9)$$

Here, category 1 has been used as a reference, and the intercept is called β_R. The other β's are regression coefficients that measure deviations from the reference category. Further, F_j, U_j and N_j are indicator variables defined as

$$F_j = \begin{cases} 1 \text{ if frequent user (category 2 and 4)} \\ 0 \text{ else} \end{cases}$$

$$U_j = \begin{cases} 1 \text{ if satisfied user (category 3 and 4)} \\ 0 \text{ else} \end{cases} \,. \qquad (10)$$

$$N_j = \begin{cases} 1 \text{ if never used ODIN or only once (category 0)} \\ 0 \text{ else} \end{cases}$$

The full model has then 10 parameters to be estimated. The model is estimated within the Bayesian framework, which means that prior distributions are specified for each parameter. Except for the inequalities (7), and the fact that all probabilities are between 0 and 1 and sum to one, no further prior information is available. Therefore we use vague priors. The prior for p_i, i=0,...,4 is determined by introducing 5 uniform variables u_i,

$$u_i \sim \mathrm{Uni}(0, 1) \qquad i = 0, \ldots, 4 \,, \qquad (11)$$

and letting p_i be given by

$$p_i = u_i / \left(\sum u_i\right) \qquad i = 0, \ldots, 4 \,. \qquad (12)$$

This induced prior was easier to handle numerically in BUGS than a Dirichlet density. Then the vector $[q_0, q_1, q_2, q_3, q_4]$ can be rewritten directly as

$$f\ [\ u_0,\ \ c_1u_1,\ \ c_2c_1u_2,\ \ c_1u_3,\ \ c_2c_1u_4] \qquad (\ 13)$$

where the factor f =$c_0/\Sigma u_i$ may be replaced by $(1/(u_0 + c_1u_1 + c_2c_1u_2 + c_1u_3 + c_2c_1u_4))$, since Σq_i=1.

The remaining prior distributions are

$$c_1 \sim \mathrm{Uni}(1, 2000) \qquad \text{and} \qquad c_2 \sim \mathrm{Uni}(1, 50), \qquad (\ 14)$$

and

$$\beta_i \sim N(0, 10^2) \qquad i\ =\ R, F, U, N. \qquad (\ 15)$$

.

The posterior distributions are estimated via the BUGS software (Spiegelhalter et. al. 1996), version 0.6 that uses the technique of Markov Chain Monte Carlo simulation (MCMC). Such methods generates a multivariate Markov Chain trajectory, one component for each of the parameters. After some burn-in (which may be long), the chain is close to the stationary distribution which is the joint posterior distribution of the parameters. Because the simulated trajectory produces dependent data, it is necessary to generate a long series to get precision. We notice that BUGS version 0.6 using the Metropolis algorithm was able to converge and deliver a fit of the model parameterized as above. With some other parameterizations the program failed.

4 Estimation results

Our main focus is on estimating the proportion of satisfied users and of frequent users among the subpopulation of those who have used ODIN more than once. This is equivalent to the probability for a person being a satisfied user or a frequent user, given that the person is in one of the categories 1 to 4. The *unweighted* proportion of frequent users defined in (1) can then also be written as

$$\text{proportion of frequent users}\ =\ (p_2 + p_4)/(p_1 + p_2 + p_3 + p_4). \qquad (\ 16)$$

The corresponding *weighted* proportion defined in (2) is given by replacing p with q. Similarly, we have

$$\text{proportion of satisfied users} = (p_3 + p_4)/(p_1 + p_2 + p_3 + p_4)\,, \qquad (17)$$

and again the corresponding *weighted* proportion is given by replacing p with q.

In order to appreciate the effect of dependency between the p's and q's, we first neglect such dependencies and estimate these proportions using each sample separately. So the p's are estimated using the telephone data only and the q's are estimated using the web data only. As noted in the introduction, this can give very misleading results for the *weighted* proportions. The results are given in Table 3. Concerning the proportion of frequent users, there is remarkable little differences between the *unweighted* (38%) and the *weighted* (41%) proportions, in light of the definition of a frequent user (daily or weekly) compared to a less frequent user (monthly or periodically or a few times). The reason may be bias due to different reasons for non-response.

Table 3 Estimated proportions of frequent users and of satisfied users, using each sample separately. 95% credibility intervals are given in parenthesis

	unweighted	weighted
proportion of frequent users	38% (23-55)	41% (33-50)
proportion of satisfied users	24% (11-40)	51% (42-60)

We now return to our model defined in last section that is based jointly on both samples. The estimated parameters are given in Table 4. The estimate of c_2 is 18.4, which is interpreted as the frequent users use in average ODIN 18.4 times more often than what less frequent users do. However, for both c_1 and c_2 the uncertainties are high. β_1 is significantly positive (but still with a high uncertainty), which is interpreted as a higher probability for non-response among frequent users than among less frequent users. Furthermore, β_2 is significant negative (with a relative small uncertainty), which means that satisfied users are more willing to answer the questionnaire, which seems to be very reasonable.

Table 4 Estimated parameters with 95% credibility intervals, based on both samples.

parameter	estimate	lower 95% cred.lim.	upper 95% cred.lim.
p_0	0.962	0.950	0.973
p_1	0.015	0.009	0.023
p_2	0.010	0.006	0.016
p_3	0.008	0.004	0.013
p_4	0.005	0.003	0.009
q_0	0.012	0.004	0.065
q_1	0.070	0.029	0.259
q_2	0.582	0.393	0.738
q_3	0.035	0.019	0.096
q_4	0.301	0.151	0.483
c_1	534	76	1590
c_2	18.4	3.0	38.4
β_R	-2.98	-11.90	2.01
β_F	7.36	1.19	18.30
β_U	-1.17	-2.20	-0.16
β_N	-4.37	-21.40	9.66

The estimated model yields Table 5, which is to be compared with Table 3. Concerning the *unweighted* proportions, the point estimate of the proportion of frequent users is unchanged, whereas the estimate of the proportion of satisfied users has increased, but is within the original credibility interval from Table 3. The uncertainty has decreased somewhat, which is natural since more information is used. It is important that the estimates of the *weighted* proportions have changed considerably. Now, the estimated *weighted* proportion of frequent users is much higher than the *unweighted* one, which is very reasonable. The estimate of the *weighted* proportion of satisfied users is much smaller in Table 5 than in Table 3, giving a less positive impression about the benefits of ODIN. Furthermore, the uncertainty is much larger in Table 5 than in Table 3, because the uncertainty in

Table 3 was calculated under the strong, and probably wrong, assumption that the probability for not answering was the same for all categories.

Table 5 Estimated proportions of frequent users and of satisfied users, by combining the samples. 95% credibility intervals are given in parenthesis.

	unweighted	weighted
proportion of frequent users	39% (27-53)	89% (64-95)
proportion of satisfied users	35% (23-47)	34% (20-52)

5 Conclusions and further research

One important conclusion from the results in the last section is that one should be very careful in interpreting results from surveys with a considerable proportion of informative non-response. This is certainly not very surprising. However, by combining data from such a survey with data from another survey with fewer data, but with better control with the type of response, it is possible to get reliable results.

The current study is based on a simple schematization of realistic surveys. We have considered only two response variables of interest, each one with only two values. With more variables, more categories and perhaps continuous, but non-Gaussian variables, the model would become much more complex and difficult to estimate. While I believe that my approach of combining surveys obtained with different sampling procedures remains valid, more research is needed to verify its practical usefulness.

Acknowledgements

I thank Arnoldo Frigessi for very helpful comments to drafts on this paper. The work was funded by the Norwegian Research Council.

References

Bachmann, D., Elfrik, J., and Vazzana, G. (1996), “Tracking the Progress of E-Mail Vs. Snail-Mail”, Marketing Research, 8, 31-35.

Solheim, I., and Tjøstheim, I. (1997), “Evaluering og brukerundersøkelse av ODIN”, NR-report 919, Norsk Regnesentral, Oslo.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996), “BUGS 0.5 - Bayesian inference Using Gibbs Sampling - Manual (version ii)”. MRC Biostatistics Unit, Cambridge.