Sampling Theory

Exam Project

Alexander James Ryan r0693898 Adityavarna Dehaleesan r0690673

14/6/2018

Contents

Introduction

When a certain population is to be judged on a characteristic, indeed there is not enough resources or time to get a response from every person. In this case, what should we do? We take a sample, a fraction of the population, from the population and we obtain responses from them. We then perform the required analysis on the information obtained and generalize it over the entire population. This has to be done effectively and for that, we would have to consider a lot of factors while selecting the sample. There will be methods that will be suitable for most of the population but might have an issue that some parts of the population cannot access the survey. For instance, if there is a survey that involves obtaining responses by e-mail or some social media, what happens to a generation of people who are older in age who do not have an e-mail id or an account, what happens to people who do not have regular access to the internet. Well in that case, it is not a very good survey because that portion of the the population's response matters as well. The more the number of people, indeed the better the result but at the same time the feasibility, human labour and time

should be taken into consideration. There are surveys that are of various kinds and an optimal solution has to be found for each of them in order to avoid inaccurate results.

In surveys of all kinds, whether a person is being interviewed in person, or whether it is a survey one had to fill up while shopping in a supermarket, or whether it is an online survey, there will be individuals or candidates who might not respond properly or not respond at all that lead to issues. In this report, we will see how we encounter such issues and whether we have any correction measures that will provide us results that are desirable.

Part One - Critique of Papers

The Problem of Non-Response in Sample Surveys

Summary

In this paper, two aspects of survey methodologies are combined as two different stages of the methodology, to reduce non response rate. The first stage being a mail questionnaire sent to a set of individuals the second step being a follow-up of personal interviews. This is also seen as a cost-effective way of survey.

Critique

The method of combining the mail questionnaire and a follow-up personal interview is a considerably good and optimal solution. This reduces the non-response and is also a very cost-effective method.

There are factors that might be affecting the exactitude of the response .Even in the cases with responses, their might be a portion of them whose response might not be the true reply. There may be many reasons for a person to not give the true response. In the mail questionnaire responses, if the questionnaire is too long, some candidates might just select options in a pattern(if it is a multiple choice questionnaire) or just write a standard reply just so that they get done with the questionnaire. This is not a good thing because this gives false information and thus leads to a faulty inference. This can be reduced by sending in short and thoughtful questionnaires so that it is good enough to not make anyone impatient about responding.

There also is an issue when the mail does not get delivered properly. So that might increase the non-response rate.

In the follow-up stage of face-to-face interview, one cannot be completely sure that the candidate will respond with complete honesty. There can be cases where he/she is not very comfortable sharing some sort of information with a stranger and might give completely dishonest replies, cases where the person responding might not give any sort of response at all, which just ends up increases the non-response rate. In such cases, the cost of conducting in-person surveys would increase but will not obtain useful data or information in return.

For the issue of a false response to the mail questionnaire, it can be made shorter and more crisp so that candidates do not get bored, lose focus and not end up answering patiently and correctly . For the case of the face-to-face interview, another follow up mail questionnaire can be sent. Telephone interviews might work better because the candidate is not physically in front of the interviewer and so the chances of being honest is more and they are also cost-effective as compared face-to-face interviews.

Conclusion

The combining of methods and using them in different stages of survey seemed to be an optimal idea. However, it cannot turn out to be as ideal as expected and there will be some issues such as the truthfulness

of the candidates' response. This can be solved to an extent by efficiently preparing the questionnaire or by modifying the survey methodology.

An Overview of Prevention and Correction Methods for Non-Response in Surveys

Summary

This article talks about prevention and correction of unit non-responses and item non-responses. A unit non-response means the person contacted refuses to to participate in the survey and an item non-response means that at the beginning the individual was willing to participate but during the survey refuses to answer certain questions. Prevention measures include shortening and crispening the questionnaire, be flexible to suit the respondent's timings for the interview and correction methods involve calibration and imputation techniques.

Critique

The article has beautifully addressed the reason for issues that can be expected while survey and has suggested a set of prevention techniques and correction methods for non-response. The issues of lengthy questionnaires that tend to tire people and the topics that are sensitive and some respondents might hesitate to share some information with the interviewer has been discussed.

In the case of person refusing to even participate in the survey, the response rate is said to increase in cases where any incentives or benefits were given to them. Interestingly, the type of interviewer also tends to have an impact of the response, for instance, a female interviewer is said to have a higher response rate, especially among the elderly sects of the population. If the reason for the non-response is something which can be helped, for example, if the person refuses to respond because he or she will not be available during the time of survey, the questionnaire dates could be changed so as to suit the respondent as well as to reduce the non-response rate. But in cases where the respondent refuses to answer because he or she does not like interviews, then there can be nothing much done from the interviewer's side. Also, in cases where the respondent cannot answer, a proxy response can be obtained. If the respondent is a young child, he or she might not be aware of the topics that are asked about. In that case, another member of the household is asked instead. Some non-responses can be avoided by choosing the right person in the household for the survey because a question about job satisfaction cannot be asked to a ten year old or an eighty year old individual. Replacement procedure is suggested if none of the above works out. These addressmost of the prominent issues that are expected to rise in a unit non-response. Now incase we get a high rate of non-response, there are some really good correction measures that are suggested. Substitution of non-respondents with individuals that have similar characteristics is a way to solve this problem. A follow-up can be conducted to those who have not responded, response rate tends to get better when asked more times. In cases where the sample is divided into classes, weights are assigned to each classes in terms of inverse of response rate. Adjusting by calibration seem to be a plausible solution because in some cases, the responses obtained from other sources are more accurate than the respective respondent answering it.

There could be many causes for the respondent to refusing to answer any particular set of questions. Individuals might not be comfortable sharing some personal or sensitive information with an unknown person. Reducing the sensitivity of the questionnaire might help reduce the error. It is interesting to note that telephone surveys in those cases tend to have a higher response rate. The topic of the questionnaire might be correlated to characteristics of the audience such as the age, gender, economic status, etc. Interviewer behaviour is a huge factor to determine the response rate of the individual. In case the interviewer is in a haste and is not clear as to what he is trying to say, it might confuse the respondent and they might end up not answering the questions. Interestingly, it is mentioned that the study shows that the difference between a study with properly formulated question, explanations and feedback, and the form with a minimum introduction is insignificant for the response rate. Indeed, Greater the effort lesser the non-response rate but we should also keep in mind

the costs involved in doing the survey. In cases of item non-response, the missing data is sometime removed completely from the analysis. It mainly uses different kind of imputations, that is, associates a particular response with a respondent based on other factors such as reporting values of neighbouring responses or replacing missing values by the average value, and rectifies the error to its best

Conclusion

The damage that non-response may cause to the survey data determined a strong reaction from the researchers in the latest years. Preventions methods are very essential because no interviewer would be interested in losing valuable data. Any survey design that aims at giving results that are as accurate as possible aims at collecting as much data within the optimal limit. The quality of the survey questionnaire, the quality of behaviour of the interviewers are all factors that influence the response rate. Prevention measures such as reducing the length of the survey or keeping in mind sensitive topic questions and correction methods such an different imputations and calibrations are used to reduce the error. Thus, prevention measures can be taken for the survey to increase the response rate and if there exists considerable non-response, a set of correction methods are essential and would be applied to the survey research to reduce the error and strengthen the validity of the results or interpretation.

Part Two - Analysis of Survey Data

Introduction

Sampling from Survey Data

Based on results found below(name of section?) the method of sampling from the Belgian population. We have proprtioned allocation based on the populations of the respective Provinces, and then devided between the 15% sample. We have taken into account the German Community, and allocated them a set 3% of the total allocation.

The method uses Simple Random Sampling, using the Floyd Ordered Hash Table Algorithm. It samples from the dataset, based on the Belgian Province variable, with a proportioned allocation.

Missing data has been removed from the dataset.

Table 1: Province Sampling

Province	Misc.	Proportion	Sample Proportion
West-Vlaanderen	-	6,72%	1,01%
Oost-Vlaanderen	-	8,38%	1,26%
Vlaams-Brabant	-	4,72%	0,71%
Antwerpen	-	10,29%	1,54%
Limburg	-	4,89%	0,73%
Brussels	-	30%	4,50%
Hainaut	-	12,18%	1,83%
Li‘ege	-	9,33%	1,40%
-	German Community	3%	0,45%
Luxembourg	-	2,53%	0,38%
Namur	-	4,44%	0,67%
Brabant-Wallon	-	3,57%	0,54%
-	Total	100,05%	15,01%

The SAS code we will use to perform this sampling strategy is thus:

To apply the proc survey select function we first had to sort the dataset by Province in ascending order.

```
proc sort data=_temp0.bmi_voeg out=work.sorted;
by province;
```

Using the proc survey select function, we chose a rate of 15%, which is asking for a 15% sample of the dataset. The strata we chose is province, and the option we selected was allocation proportional, which tells the function to sample from the dataset, based on the proportion of the population in each Province.

```
title1 'Belgian Health Survey';
title2 'Stratified Sampling of West-Vlaanderen';
proc surveyselect data=work.sorted
rate=0.15 out=work.SampleControl;
strata PROVINCE / ALLOC=PROP;
run;

proc means data=work.samplecontrol;
run;
```

Table 2: Descriptive Statistics for BMI

Basic Statistical Measures	-	—-	—-_1
Location	Variability	-	-
Mean	2.445.393	Std Deviation	430.323
Median	2.387.543	Variance	1.851.776
Mode	2.249.135	Range	3.278.068
-	-	Interquartile Range	508.875

[ADD descriptive statistics for Province]

Modelling BMI

We chose to remove records with missing values. First we devided the variable BMI into an ordinal categorical variable. Then a ordinal logistic regression model was built. We used a stepwise iterative procedure to incrementally add variables, with AIC minimisation the target goal. The model we are left with is:

With a AIC score of 12973.46 and a residual deviance of 12939.46.

Table 3: Ordinal Logistic Model 1

X1	Value	Std. Error	t value
AGE72	113.32800	0.08334	135.9780
AGE73	154.07600	0.08490	181.4830
AGE74	202.04100	0.09135	221.1670
AGE75	227.64000	0.09661	235.6380
AGE76	217.39300	0.09758	222.7820
AGE77	140.95900	0.11047	127.6020
SEX2	-0.72724	0.04887	-148.8070
EDU32	-0.23508	0.06035	-38.9500
EDU33	-0.52241	0.06392	-81.7290
SGP1	0.30198	0.08015	37.6770
TA22	-0.18705	0.04882	-38.3150
WAL1	0.15514	0.04896	31.6860

X1	Value	Std. Error	t value
INCOMELOW	0.16401	0.07684	21.3440
INCOMEMED	0.04015	0.07550	0.5318

Coefficients from are difficult to interpret since they are scaled in terms of logs. To interpret the results of a logistic regression model we can convert the coefficients into an odds ratio. To do this, we just exponentiate the estimates and confidence intervals.

Interpretation

For the variable SEX, being a Man, i.e., going from 0 (Woman) to 1 (Man), the odds of "BMI<20" versus "2025" combined are 0.48 greater, ceteris paribus.

[more interpretation. . .]

Part Three - Design of Health Interview Survey

Introduction

The choice on where to stratify and group populations is difficult. The goal is to create a sample that is representative of different levels of the population, so that at each level of analysis, we can form accurate estimands. Should the focus be on weighting regions, then perhaps some large cities in smaller regions become under-represented, and small cities in large regions become over-represented. Ultimately the strategy in which you choose to stratify a population, will over-represent some communities, and under-represent others, an implicit bias that is impossible to evade. The goal of sampling populations is to reduce this imbalance as much as possible, so every community is given a fairly weighted representation.

In this paper, we will discuss a variety of stratifying strategies, to illustrate the complexity and variation in choosing sampling priorities. The first division is in the three different regions in Belgium, Walloon, Brussels and Flanders.

Table 4: Proportioned weights at Region level

Region	Weight
Wallonia	35%
Brussels	30%
Flanders	35%

Given that we have to exclude 3% of our total 35% sample of the Walloon Region for the Germany community, it is important how we allocate the remaining 32% amongst the remaining communities. Depending on how we allocate, we could bias the results, over representing some regions, whilst under-representing others.

For example, if we simply divided the 35% total allocation for the Walloon Region, and divided by 5, we would be sampling 7% people from each region. Since for the Liege province, we have set aside 3% of that for the German community, we only have 4% left for the Arrondissements Huy (pop=189,661), Liege (pop=617,551), Verviers (pop=208,249). Liege is the largest Arrondissement in the Walloon region, and we would find it severely under-represented if we use this stratifying strategy.

Stratifying Strategy 1

Each Region weighting is devided equally between Provinces and Arrondissements

Of the 35% allocated to Flemish Region, each of the different Provinces receives 7%. This 7% is then proportionally allocated to each Care Sector based on their respective populations. The same strategy is then used for the Walloon Region, but first the set allocation of 3% to the German community is taken from 35%, leaving 32% to be allocated to the Walloon Provinces Thus each Walloon Province receives 6.4%, which is then proportionally allocated to each Arrondissement based on their respective populations.

Table 5: Stratifying Strategy 1

Province	Care sector/Arrondissement	Flemish Region	Brussels	Walloon Region
West-Vlaanderen	Brugge	1,87%	-	-
-	Oostende	1,12%	-	-
-	Roeselare	2,13%	-	-
-	Kotrijk	1,88%	-	-
Oost-Vlaanderen	Gent	4,23%	-	-
-	Sint-Niklaas	1,14%	-	-
-	Aalst	1,63%	-	-
Vlaams-Brabant	West	2,90%	-	-
-	Leuven	4,10%	-	-
Antwerpen	Antwerpen	3,68%	-	-
-	Mechelen	1,59%	-	-
-	Turhout	1,74%	-	-
Limburg	Hasselt	4,93%	-	-
-	Genk	2,07%	-	-
Brussels	-	-	30%	-
Hainaut	Tournai	-	-	1,48%
-	Charleroi	-	-	2,06%
-	Thuin	-	-	0,73%
-	Mons	-	-	1,23%
-	Soignies	-	-	0,90%
Li‘ege	Huy	-	-	1,20%
-	Li‘ege	-	-	3,89%
-	Verviers	-	-	1,31%
-	German Community	-	-	3%
Luxembourg	Arlon	-	-	3,75%
-	Neufchateau	-	-	2,65%
Namur	Namur	-	-	4,09%
-	Dinant	-	-	2,31%
Brabant-Wallon	Nivelles	-	-	6,40%

Stratifying Strategy 2

The 35% allocated to the Flemish Region is proportionally allocated to each Care Sector, based the proportion of their populations in respect to the total population of the Province. The same strategy is used for the Walloon region, and again, removing the 3% allocated to the German Community, leaving 32% to be proportionally allocated to the Arrondissements.

Table 6: Stratifying Strategy 2

Province	Care sector/Arrondissement	Flemish Region	Brussels	Walloon Region
West-Vlaanderen	Brugge	1,80%	-	-
-	Oostende	1,08%	-	-
-	Roeselare	2,04%	-	-
-	Kotrijk	1,80%	-	-
Oost-Vlaanderen	Gent	5,07%	-	-
-	Sint-Niklaas	1,36%	-	-
-	Aalst	1,95%	-	-
Vlaams-Brabant	West	1,95%	-	-
-	Leuven	2,77%	-	-
Antwerpen	Antwerpen	5,41%	-	-
-	Mechelen	2,33%	-	-
-	Turhout	2,55%	-	-
Limburg	Hasselt	3,45%	-	-
-	Genk	1,44%	-	-
Brussels	-	-	30%	-
Hainaut	Tournai	-	-	2,81%
-	Charleroi	-	-	3,93%
-	Thuin	-	-	1,38%
-	Mons	-	-	2,35%
-	Soignies	-	-	1,71%
Li‘ege	Huy	-	-	1,73%
-	Li‘ege	-	-	5,65%
-	Verviers	-	-	1,90%
-	German Community	-	-	3%
Luxembourg	Arlon	-	-	1,48%
-	Neufchateau	-	-	1,05%
Namur	Namur	-	-	2,84%
-	Dinant	-	-	1,60%
Brabant-Wallon	Nivelles	-	-	3,57%

Stratifying Strategy 3

Using a combination of the two strategies above, the Provinces receive a weight based on respective population size of the total Region population. Then another weight is calculated, to proportion the Care Sectors/Arrondissements based on their population to their respective Province. The German community was allocated 3%, and removed from the stratifying strategy, so the Walloon region, excluding the German Community, was allocated 32% in total.

Table 7: Stratifying Strategy 3

Province	Care sector/Arrondissement	Flemish Region	Brussels	Walloon Region
West-Vlaanderen	Brugge	1,80%	-	-
-	Oostende	1,08%	-	-
-	Roeselare	2,04%	-	-
-	Kotrijk	1,80%	-	-
Oost-Vlaanderen	Gent	5,07%	-	-
-	Sint-Niklaas	1,36%	-	-
-	Aalst	1,95%	-	-
Vlaams-Brabant	West	1,95%	-	-

Province	Care sector/Arrondissement	Flemish Region	Brussels	Walloon Region
-	Leuven	2,77%	-	-
Antwerpen	Antwerpen	5,41%	-	-
-	Mechelen	2,33%	-	-
-	Turhout	2,55%	-	-
Limburg	Hasselt	3,45%	-	-
-	Genk	1,44%	-	-
Brussels	-	-	30%	-
Hainaut	Tournai	-	-	2,81%
-	Charleroi	-	-	3,93%
-	Thuin	-	-	1,38%
-	Mons	-	-	2,35%
-	Soignies	-	-	1,71%
Li‘ege	Huy	-	-	1,73%
-	Li‘ege	-	-	5,65%
-	Verviers	-	-	1,95%
-	German Community	-	-	3%
Luxembourg	Arlon	-	-	1,48%
-	Neufchateau	-	-	1,05%
Namur	Namur	-	-	2,84%
-	Dinant	-	-	1,60%
Brabant-Wallon	Nivelles	-	-	3,57%

Part Four - Missing Data in the Belgian Health Interview Survey

Introduction