CENTRE FOR
COMPUTING AND BIOMETRICS

# Surveying a Community's Perception of Odour

Jim Young

Research Report No:97/02
February 1997

RESEARCH REPORT

ISSN 1173-8405

LINCOLN
UNIVERSITY
Te Whare Wānaka O Aoraki

# Centre for Computing and Biometrics

The Centre for Computing and Biometrics (CCB) has both an academic (teaching and research) role and a computer services role. The academic section teaches subjects leading to a Bachelor of Applied Computing degree and a computing major in the BCM degree. In addition it contributes computing, statistics and mathematics subjects to a wide range of other Lincoln University degrees. The CCB is also strongly involved in postgraduate teaching leading to honours, masters and PhD degrees. The department has active research interests in modelling and simulation, applied statistics and statistical consulting, end user computing, computer assisted learning, networking, geometric modelling, visualisation, databases and information sharing.

The Computer Services section provides and supports the computer facilities used throughout Lincoln University for teaching, research and administration. It is also responsible for the telecommunications services of the University.

## Research Report Editors



## Copyright



## Correspondence

This paper represents work to date and may not necessarily form the basis for the authors' final conclusions relating to this topic. It is likely, however, that the paper will appear in some form in a journal or in conference proceedings in the near future. The authors would be pleased to receive correspondence in connection with any of the issues raised in this paper. Please contact the authors either by email or by writing to the address below.

Any correspondence concerning the series should be sent to:

The Editor
Centre for Computing and Biometrics
PO Box 84
Lincoln University
Canterbury, NEW ZEALAND

Email: computing@lincoln.ac.nz

# Surveying a community's perception of odour

**Jim Young**
**Centre for Computing and Biometrics**
**Lincoln University, New Zealand**
**young2@lincoln.ac.nz**

## 0. Foreword

I wrote this document as part of a research assignment funded by the Ministry for the Environment and by a number of local authorities and private companies. The assignment, 'Guidelines for community odour assessment', was to prepare advice for those planning to collect information on a community's perception of odour. My job was to write about the statistical aspects of surveying a community. This was only a small part of the assignment: Carolyn Blackford did the rest - including a review of case law, a review of overseas practice, advice on questionnaire design.

Although this is about surveying a community for its perception of odour, most of what is in this document will apply to any survey of a community. The document appears as I first wrote it, before it was incorporated into a final report. I have added this foreword and an appendix. The appendix expands on comments made at the end of section 4. I wanted to summarise how I thought it easiest to survey individuals when an over-estimate of the problem would be an advantage. But I thought this material would be more difficult than what is in the rest of this document, and so I made no attempt to write this material up in time to include it in the final report.

I am most grateful to those who funded this assignment and to Carolyn Blackford, and to those at Statistics New Zealand who taught me the tricks of the trade - especially Mike Doherty, Alistair Gray and Richard Penny.

# 1. Introduction - how to use this document

To use this document, you will need to know some basic statistics - a good pass in sixth form statistics or mathematics should be enough. You should first read sections 2 to 5. These sections are about survey objectives, defining a population of interest, how to represent a community's views, and some general principles of sample design. Once you have read these sections, you will know which of the specific sample designs (sections 6 to 9) is going to be of most use. But it's probably better to read on until you come to the design you want, because the methods used in sections 8 and 9 are explained in more detail in sections 6 and 7.

Section 2 covers the different objectives councils and odour producers may have when surveying a community. While a council could also be an odour producer, in general there will be three parties: a council, an odour producer, and the rest of the community. Both councils and odour producers need to select a sample of people to represent the community. A council often has a list of ratepayers from which it can draw its sample. An odour producer will typically not have this sort of information. So, councils and odour producers will tend to use different sorts of surveys because their objectives differ, and because councils already know where people live and work in the community.

Section 3 is about how to define a population of interest. The population of interest is the group of people whose opinions you want to record. Obviously, if you define your population as those living within a certain area, and this area is very large, then the proportion of people who perceive an odour problem may be very small. So this definition of your population is crucial to the subsequent use you can make of your survey's results.

Section 4 is about how to represent the community's views. While it is relatively easy to sample households and businesses in an area, sampling individuals within each household or business adds another layer of complexity to the sampling process. You need to consider whether this added complexity is necessary.

Section 5 is about general principles of sample design - the statistical aspects of running a survey. Four major steps in the sample design process are: developing a sampling frame (a list of those in the population of interest), selecting people for the survey, estimation (calculating results once the survey is completed) and adjusting results for non-response. Sample design takes place in conjunction with questionnaire development and survey management planning. Each of these three operations affects the others. Statistics New Zealand's 'A guide to good survey design' is full of good practical advice on running a survey.[1]

## 2. Survey objectives

Councils and odour producers may survey a community's perceptions of odour for different reasons. While a council could also be an odour producer, in general there will be three parties: the council, whose job it is to monitor odour and complaints about odour and to consider resource consent applications for permission to discharge odour; the odour producer - that part of the community producing the odour; and the community - the rest of the community, those not responsible for the odour but actually or potentially affected by it.

Surveying a community's perceptions of odour is not the same as environmental monitoring. The latter is concerned with whether there is odour; the former is concerned with whether odour is perceived to be a problem. Environmental monitoring will most likely involve spatial sampling, rather than survey sampling.[2]

If there is more than one source of odour, a community survey may not tell you which source is causing a problem in the community. Collecting data from people (perhaps using odour diaries) on the meteorological conditions at the time they detect an odour will help. But to really separate the effect of each individual source will probably require environmental monitoring and chemical analysis.[3]

In surveying the community, the council or odour producer could be seeking information for its own internal use, or for legal purposes. I will call the first sort of survey an 'indicative' survey, and the second sort a 'definitive' survey. Indicative surveys need to be quick and cheap. Definitive surveys need to be able to withstand legal challenge. Criteria to be met before survey results will be admitted as evidence are discussed in *Auckland Regional Council v Mutual Rental Cars*.[4]

The council may wish to run an indicative survey to determine whether a complaint represents a wider problem (which will then need to be acted on), or whether the complaint is 'vexatious'. The odour producer may wish to run an indicative survey to gather information which will help it to better manage odour. Both council and odour producer may wish to run definitive surveys as part of the resource consent process, or when enforcement orders or abatement notices are called into question.

Some sort of sampling frame is necessary in any survey. A sampling frame is a list of those in the population of interest. A district or city council will be able to use its list of ratepayers. A regional council or an odour producer may have to use some other sort of list. In what follows, I will refer to a 'council' as though all councils can use a list of ratepayers as a sampling frame, and I will refer to an 'odour producer' as though all odour producers cannot use a list of this sort. This is not true and so a council may find itself using guidelines more commonly used by an odour producer and the other way around.

So with these definitions, each combination of 'council' or 'odour producer' and 'indicative' or 'definitive survey' will require a different sort of sample design.

## 3. Defining a population of interest

The starting point for any survey is a definition of the population of interest. The definition should be clearly stated. The definition will most likely include statements about an area in which the community to be surveyed lives, and about a time period of interest. The aim is to define an area and a time period within which a community is likely to have found odour offensive.

Those planning a definitive survey must take care that the definition used can withstand legal challenge. For example, if odour quickly dissipates with distance from its source, increasing the area of interest will reduce the proportion in the survey who perceive a problem. A survey at the end of the summer is not a sensible way to assess an odour problem that occurs only in winter.

To decide on a population of interest, and to later defend that decision, you could use: addresses and dates of complaints received; results of atmospheric dispersion modelling;[5] documented visits by you or others to different areas; populations of interest used elsewhere for similar problems; discussion of how local climate and topography influence odour dispersion. Much of this information can be displayed using maps.

The Ministry for the Environment suggests that 'an offensive odour generally becomes noticeable when its concentration reaches 5-10 OU' (odour units).[6] So atmospheric dispersion modelling, which gives contours in odour units, could be used to set a boundary to an area of interest. However, since models generally assume constant climatic conditions and crudely approximate topographical features, the addresses of those complaining will be the better guide.[5]

For a definitive survey, it may be necessary to survey outside your population of interest as well. This is to demonstrate that you have done a good job in defining those likely to find odour a nuisance.[7] However, you can use different sampling fractions in and outside the population of interest, sampling fewer outside the population of interest in order to keep the total sample size down.

*3.1.1 Given the addresses of complaints you've received, you define a population of interest as those living or working inside a circle with a radius of 1.0 km. At the centre of the circle lies a factory you suspect is causing an odour problem. To show that few people outside this area find odour offensive, you also survey those living between 1-2 km away from the centre of the circle.*

## 4. Representing a community's views

A community consists of those living and working in an area, those visiting an area and those investing in an area. But individuals in these last two categories will be hard to identify. In most cases, it should be enough survey just those living and working in an area, and make some comment about how these other two groups could be affected in the light of your survey's results. In special circumstances (perhaps a large park, tourist attraction, or shopping centre), surveying these other two groups may be necessary. These guidelines do not address how to survey visitors nor investors.

So you will need to sample both businesses and households. It is relatively easy to tell how many businesses and households are in an area. You might use a list of ratepayers, a list of telephone numbers, aerial photographs or electoral role information [the 'Habitation Index']. And you can then sample some of these businesses and households. But you need to talk to individuals, and so you need some way of choosing which person or people to interview within each of the businesses or households in your sample. This means, in sampling terms, a second stage of selection: the first stage is to sample households and businesses; the second stage, to sample individuals from within the previously selected households and businesses. If possible, you want to avoid this second stage of selection for two reasons.

Firstly, you don't know how many people are in each primary sampling unit (the business or household). So the interviewer has to find this out, and then choose at random a person to interview. This means more (and difficult) work for the interviewer, and interviewers will need more training as a result.

Secondly, estimates from a survey are not enough. You need to show how precise your estimates are. To do this, you need to calculate the variance in what you are trying to measure. With two stages of sampling, both stages contribute to the variance.[8] But to estimate the contribution to the variance from the second stage, you need to interview more than one person in each household. This is difficult to do without the answers of the first person interviewed influencing the answers of the second person interviewed.[9]

One way around the problem is to always interview 'the household member who is most knowledgeable'.[10] I suggest 'the person most often at home' should be the most knowledgeable member of the household on matters of offensive odours, and a Health and Safety Officer (if on site) should be the most knowledgeable employee in a business. That is, if the business has such a person on site: you would have to approach the CEO first, and ask to speak to the most appropriate person (the person who would receive any complaints about odour from those employed by the business).

However this solution means that the results of your survey will be in terms of households (and businesses), not individuals. That is, '15% of households within 500 metres of the source found the odour offensive', not '15% of the individuals living within 500 metres'. And you will need to explain your procedure for choosing someone to represent the views of each household or business (particularly if it is a definitive survey).

Selecting individuals this way has some real practical advantages. Firstly 'the person most often at home' will be available for an interview more often than someone chosen at random from those in the household. This will reduce the number of households you have to visit (or phone) a second time in order to interview the person you want. Secondly the survey may be repeated some time in the future, if estimates of change are required. It is more accurate to use the same sample (both 'before' and 'after') so that the change is not partly due to a difference between two different samples. If the same sample is used, a sample of differences

('after' minus 'before') is then used to estimate the average difference in the population. Finding 'the person most often at home' in a second survey (and this person could legitimately change between surveys) is much easier than finding some specific individual, chosen at random in the first survey.

On the other hand, a community is really made up of individuals, not households and businesses. 'The person most often at home' is more likely to find an odour offensive than someone chosen at random from the household. So using the 'person most often at home' approach will over-estimate how many individuals perceive a problem. Of course, as an odour problem becomes more and more obvious, this bias disappears because it won't matter who you talk to in the household - everyone will tell you there's a problem!

If you are running an indicative survey, a slight over-estimate won't really matter. With a small sample, the bias (the difference, on average, between what the survey tells you and the true value) will be small compared to the variance in the estimate. When the variance is large, this is saying the estimate isn't known precisely anyway. Selecting individuals this way will actually be an advantage to the odour producer who wishes to gather information in order to better manage odour. In this case, you want to talk to those who think there is a problem.

In a definitive survey, an over-estimate of the individual's views will not be a problem if it does not help your cause. For example, the odour producer who wishes to show there is a low level of community concern will not benefit from using results from 'the person most often at home' as representing the level of individual concern.

However, if a definitive survey of individuals is necessary and an over-estimate is going to be to your advantage, then you may need to randomly sample individuals within households and businesses. At this point, consult a professional. Briefly, the common method of selecting the individual who has the next birthday is not recommended.[11] Kish gives a better method: the interviewer ranks household members by increasing age (the interviewer doesn't actually need to know these ages), then uses a random number table to choose one person from this list.[9] Estimates need to be adjusted to take account of the variable selection probabilities. The variance in these estimates can be calculated as if there were no second stage of sampling. This will under-estimate the true variance,[12] but Kish suggests the difference will not be large provided most households and businesses are about the same size (say one to six people).[13] An appreciable number of larger businesses in the population of interest means that businesses should be treated differently. A sample of people will need to be selected from each business, and second stage variances calculated for the business component of the community. Alternatively, businesses could be sampled with replacement (see section 5.2) and then there is no need to calculate second stage variances.[14]

# 5. General principles of sample design

## *5.1 Sampling frame*

The sampling frame is a list of all households and businesses available for selection. If you are lucky, this list will consist of every household and business in your population of interest. Often there is some mismatch between the sampling frame and the population of interest. For example, a household without a phone or with an unlisted phone number will not be in a sampling frame compiled from the telephone directory. Duplicates in the sampling frame will cause results to be biased towards the opinions of the duplicated group. It is good practice to report any mismatch between sampling frame and the population of interest and discuss how this could affect results. As a result of this mismatch, do you expect estimates to be too high or too low for certain questions?

## *5.2 Selecting the sample*

A sample is then selected from the sampling frame. Random sampling ensures that one can estimate (without bias) the properties of the population of interest from the sample. In a simple random sample, each item in the sampling frame has an equal chance of being selected for the survey. Other sorts of random sampling are more efficient than simple random sampling. That is, for a given sample size, other sorts of random sampling can give more precise estimates (precision is measured by the variance of the estimate - see section 5.3). So before selecting the sample, you need to calculate a suitable sample size but this calculation depends on the sort of random sampling you're going to use. Appropriate methods of random sampling will be introduced as needed, together with how to calculate a suitable sample size.

The simple random sample (without replacement) is a component in many of the more complicated methods of random sampling. Sampling with replacement means that when an household or business is selected from the population for the sample, it is observed and then returned to the population before the next draw. Sampling without replacement ensures that once an household or business is selected for the sample, it cannot be selected a second time. Simple random sampling is usually without replacement because it is more efficient.[15]

The easiest way to take a simple random sample (without replacement) is to use a computer program (such as a spreadsheet or statistical package) to select your sample. You could read your sampling frame into the computer program and then select the sample. Alternatively, you could create a column of numbers inside the computer program, so that each number represent one item in your sampling frame; then select numbers from this column and use these numbers to identify the items in your sampling frame that are in the sample. Most computer programs will have simple random sampling without replacement as the default, but it pays to check.

If you can't use a computer to draw the sample, you can take a systematic sample from your sampling frame by hand. To draw a systematic sample of roughly size n, choose a random number between one and k, and then select every $k^{th}$ ratepayer from the sampling frame starting with the ratepayer corresponding to the random number. Calculate k as the integer part of N/n, where N is the total number of ratepayers in the sampling frame.

5.2.1 *You calculate the required sample size to be 60. There are 893 ratepayers in the sampling frame, so k = 893 / 60 = 14.88 ⇒ 14. You now need to find a random number between 1 and 14, as a starting point for sampling from the list. You get a scientific calculator to choose a random number between 0 and 1: the calculator selects 0.271. Multiply this number by k; add one; and then just use the integer part of the answer: (0.271×14) +1 = 4.794⇒ 4. So select the fourth ratepayer in the list, and then every 14th ratepayer after that, until you get to the end of the list.*

A systematic sample from a list arranged alphabetically can be treated as if it were a simple random sample.[16] If the list is not in alphabetic order, make sure the order is essentially random. Particularly watch out for any order in the list that roughly coincides with the frequency of sampling from the list.[17]

5.2.2 *In the example above (section 5.2.1), suppose the sampling frame consisted mainly of households but every 14th ratepayer in the sampling frame was the owner of a business. The systematic sample would then either have no businesses in it at all or would consist entirely of businesses, depending on the random starting point chosen. While an exact periodic effect like this is unlikely, roughly periodic variation in the list is a possibility and this could lead to biased estimates. If in doubt, take a simple random sample instead.*

### 5.3 Estimation

With a simple random sample, the sample average is an unbiased estimate of the population average. This is not true for most other random sampling methods. Once results are in, estimates of population averages or totals must be calculated. Mostly, estimate of averages will be required: 'on the last occasion odour was perceived as annoying, the odour lasted on average for 1.8 hours'. Note that a proportion is a special sort of average. Proportions are converted to percentages by multiplying the proportion by 100. So '15% of respondents felt the odour was offensive' is a statement about a proportion. If each person who feels odour is offensive is recorded as a one, and a zero is recorded otherwise, adding up the ones and zeros and dividing by the number of respondents yields a proportion. Yet this process is just finding an average.

The likely variation in estimates must also be calculated. This 'variance' is a measure of how precisely answers are known. In statistics, precision refers to how much an estimates would vary about an average value if you repeatedly sampled your population - but if the estimate is biased, this average value will not be the true value. In general, one can calculate the precision of an estimate but not its bias.

For legal purposes, knowing the precision of the estimate is as important as knowing the estimate itself. If a survey is to be used as evidence, details may be required of 'any tests applied and the results of any tests applied to determine the extent to which the survey or results of the survey can be trusted'.[4]

One appropriate test is the 95% confidence interval. This interval is approximately the estimate (of an average, total or proportion) plus or minus two times the square root of the variance of the estimate.

$$\text{ie. 95\% CI} = \text{estimate} \pm 2.\sqrt{\text{variance of estimate}}\ . \qquad 1.$$

The interpretation of this interval is that for 95% of all possible samples, the true value for the population of interest will lie within the interval. Hence the true value is to be found within

this interval with a high degree of confidence. As a general rule, to be reliable a confidence interval needs to be based on a sample size of at least 30.[18,19]

5.3.1 *You take a simple random sample of 30 households from a population of 2000. In five of these 30 households, the person 'most often at home' says that odour is a problem. Therefore the proportion of households where odour is perceived as a problem is 5/30 or 0.17 (17%). Let's say the variance of this proportion is 0.0047 (you'd use equation 11 in section 6.3). The 95% confidence interval is then 0.17 ± (2×0.07), or from 0.03 to 0.31. That is, in 3% to 31% of households in the population of interest, the 'person most often at home' says that odour is a problem.*

Other higher or lower percentage confidence intervals could be useful. Most statistics textbooks give examples of how to do this. The 95% confidence interval has a long history of use in science, and should be acceptable for legal purposes. But there's nothing magic about this 95% confidence interval - for indicative surveys, 90% or 80% confidence intervals may well be more useful. Indicative surveys are likely to use small samples. The result will be 95% confidence intervals that are very wide: perhaps so wide as to be rather uninformative (as in section 5.3.1 above). The 80% interval will give you a narrower interval within which the true value is likely to lie, although you cannot be quite so certain that the true value is within this interval. The 80% confidence interval is approximately the estimate (of an average, total or proportion) plus or minus 1.3 times the square root of the variance of the estimate:

$$\text{ie. } 80\%\ \text{CI} = \text{estimate} \pm 1.3.\sqrt{\text{variance of estimate}}\,. \qquad 2.$$

### 5.4 Non-response

Finally, those who do not respond to a survey are often different to those who do. Non-response introduces a bias: on average estimates from the survey differ from the true values for the population. Many things can be done, as part of survey management, to reduce non-response. With a definitive survey, you should try to contact those not at home at least three times before giving up. With an indicative survey, try to contact those not at home at least twice (or at least try some them a second time - see section 6.4). Other ways to reduce non-response include contacting people in advance to arrange an interview time, assurances of privacy and confidentiality, providing some small incentive.[20] Good questionnaire design helps too. But there will always be some degree of non-response. Once answers are in, statistical methods can be used to adjust estimates so they are less biased.

You should always report your survey's response rate. As a rule of thumb, the response to your survey should be 70% or better. Otherwise 'most researchers would have more accurate and useful estimates if they reduced the sample size and devoted the saved resources to obtaining responses from a higher percentage of the sample'.[21]

# 6. Council: indicative survey

## *6.1 Sampling frame*

A list of ratepayers (both households and businesses) in the area of interest will be the best sampling frame. Those selected from the list will be contacted, and the person most often at home will be asked a short series of simple questions.

A telephone survey will be the quickest and cheapest way to reach those selected.[22] Postal surveys are not recommended because of their low response rates.[23] A telephone survey needs a short questionnaire and simple questions. A short questionnaire is a good selling point when you're trying to persuade someone to participate in your survey. Simple questions are needed for telephone interviews; otherwise the respondent may have difficulty understanding what is being asked. [In face to face interviews, it's easier to tell if the respondent doesn't understand the question and needs additional information from the interviewer. The respondent can also see a copy of a long question, or the choice of answers as the question is being asked by the interviewer.]

Treat flats as businesses. In the first instance, interview the owner (as the CEO): he or she may then suggest you speak to a tenant (as a more appropriate person to answer your questions) and may help you contact this person. For an indicative survey, it won't really matter if you have to interview the owner when you'd rather interview a tenant (see comments in section 4 about bias versus variance in estimates from indicative surveys).

## *6.2 Selecting the sample*

Take a simple random sample (or systematic sample) from your list of ratepayers (section 5.2). To calculate the required sample size, identify the most important question or questions in your survey. For each question, consider the width of the confidence interval you want to end up with. Calculate an approximate sample size for each question as:[24]

$$n_0 = \frac{Z^2\sigma^2}{\partial^2},$$

$n_0$ = first approximation to required sample size;

$Z = 2.0$ for a 95% confidence interval; 3.

$= 1.3$ for an 80% confidence interval;

$\partial$ = half the width of the desired confidence interval;

$\sigma^2$ = population variance (as yet unknown).

Now adjust this approximate sample size given the number in the sampling frame (N):[25, 26]

$$n = \frac{n_0 N}{n_0 + N},$$

n = required sample size; 4.

N = population size (number in the sampling frame).

Do this for each important question and take the largest value of n as the sample size required for your survey. Remember, you need a sample size of at least 30 to be able to calculate reliable confidence intervals (section 5.3).

Of course you don't know the population variance before surveying, and the trick is to make an intelligent guess. If you are going to estimate an average, think of the largest and smallest values you are likely to get in answer to your question. The population variance is then roughly:[27]

$$\sigma^2 = \left[\frac{(\text{largest} - \text{smallest})}{6}\right]^2. \qquad 5.$$

*6.2.1 You are going to ask respondents: 'Think back to the last time you were annoyed by odour. How long did the odour last?'. You want to estimate how long offensive odours last. You think the answers you'll receive will range from 0 to 15 hours. Your rough estimate of the population variance is then $[(15-0)/6]^2=6.25$. If you decide you want to calculate an 80% confidence interval for this mean with width plus or minus half an hour, your initial sample size is $1.3^2 \times 6.25 / 0.5^2 = 42$. If there are 2000 ratepayers in your sampling frame, the required sample size is adjusted to $(42 \times 2000) / (42+2000) = 41$. Select say 60 people to allow for a 70% response rate. Note that in this example, using equation 4 doesn't really change the sample size because here the sample size is small relative to the population size.*

If you are going to estimate a proportion, note that the population variance for a variable that can only take the values zero or one is approximately:[28]

$$\sigma^2 = P(1-P). \qquad 6.$$

This equation is at a maximum when the population proportion (P) is 0.5 (that is, 50%). So think of what the population proportion is likely to be; then choose a value closer to 0.5. Without any information at all, you might use P = 0.5; but you risk being too conservative and ending up with a much larger sample size than you really need.

*6.2.2 You think around 10% of the population will find a particular odour offensive. You want a 95% confidence interval with width plus or minus 10%. [That is, if the estimate of this proportion turns out to be 0.12, you want to end up with a 95% confidence interval from 0.02 to 0.22.] To be slightly conservative, you assume that the population proportion (P) is 0.15. Your initial sample size is $2.0^2 \times 0.15 \times 0.85 / 0.1^2 = 51$. Notice that if you'd assumed P = 0.5, your initial sample size would be 100. If there are 200 ratepayers in your sampling frame, the required sample size is adjusted to $51 \times 200 / (51+200) = 41$ - say 60 people to allow for a 70% response rate.*

### 6.3 Estimation

Once the survey has been completed, estimate an average as:[29]

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n},$$

$\hat{\bar{Y}}$ = estimate of population average;

$\bar{y}$ = sample average; 7.

$\sum_{i=1}^{n} y_i$ = add up each sample observation;

n = sample size.

The variance of this estimate is:[30]

$$v(\bar{y}) = \frac{(N-n)}{N} \cdot \frac{s^2}{n},$$

$v(\bar{y})$ = variance of sample average; 8.

$N$ = population size;

$s^2$ = sample variance.

Use the statistical function of a scientific calculator to find the sample variance ($s^2$), or calculate this by hand as:

$$s^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1},$$

$\sum_{i=1}^{n} y_i^2$ = add up the square of each sample observation; 9.

$\left(\sum_{i-1}^{n} y_i\right)^2$ = add up each sample observation, then square the result.

*6.3.1 Each person in a sample of 40 tells you how long (in hours) the odour lasted in their most recent experience of an offensive odour (see section 6.2.1). The sum of each sample observation is 72, and the sum of each sample observation squared is 320. The variance is therefore [320-(72²/40)]/39 = 4.88. The mean is 72/40 = 1.8 hours, the variance of the mean (equation 8) is [(2000-40)/2000]×(4.88/40) = 0.12. From section 5.3, an 80% confidence interval for this mean is 1.8 ± (1.3×0.35) - that is 1.3 to 2.3 hours.*

Estimate a proportion (see section 5.3) and its variance as:[31]

$$p = \frac{a}{n},$$

$p$ = sample proportion (see Section 5.3); 10.

$a$ = number in sample with value one rather than zero;

$n$ = sample size.

The variance of this proportion is:[32]

$$v(p) = \frac{(N-n)}{N} \cdot \frac{p(1-p)}{(n-1)},$$

$v(p)$ = variance of sample proportion; 11.

$N$ = population size.

If you then calculate a confidence interval (section 5.3) and find that zero or one is included in this interval, you will need to use a more accurate method.[33]

6.3.2 *You ask a sample of 40 people (from the 200 in your sampling frame - see section 6.2.2) if they have experienced any offensive odours in the last week. Of the 40, 5 answer 'yes'. Your estimate of the proportion is 5/40 = 0.125 and the variance of this proportion is [(200-40)/200]×(0.125×0.875/39) = 0.0022. From section 5.3, a 95% confidence interval for the proportion experiencing offensive odours is 0.125 ± (2×0.047) = [0.03, 0.22] - that is, 3 to 22%.*

## 6.4 Non-response

If you have the resources, you should make at least two attempts to contact 'the person most often at home'. One way to save resources in this process is follow up just a simple random sample of those who weren't contacted in the first attempt.

You take a sample of n people: you have contacted $n_1$ of these people but you haven't been able to reach $n_2$ ($n_1 + n_2 = n$). Take a simple random sample (or systematic sample - see section 6.2) of size r from the $n_2$ people you haven't contacted. Make a major effort to get responses from these r people. Instead of using the sample average, estimate the population average as:

$$\bar{y}' = \frac{(n_1\bar{y}_1 + n_2\bar{y}_r)}{n},$$

$\bar{y}'$ = adjusted sample average;

$\bar{y}_1$ = average for those contacted initially; 12.

$n_1$ = number contacted initially;

$n_2 = n - n_1$;

$\bar{y}_r$ = average for the r respondents contacted subsequently.

The difference between this estimate and the normal sample average should be reported. This difference is a measure of the bias that results from being unable to contact some people. Note that since a proportion is just a special sort of average (section 5.3), you can replace averages in the above equation with the appropriate proportions.

6.4.1 *You carry out the survey described in section 6.2.2. The response rate in your survey is lower than the 70% you hoped for. Of the 60 ratepayers in your sample, you manage to contact only 35 on your first attempt. Of these 35, 4 say they have experienced offensive odours in the last week. From the 25 non-respondents you choose 10 at random, and pursue these people until at last you make contact. Of these 10, 3 answer 'yes' to your question. Your proportion adjusted for non-response is [35×(4/35) + 25×(3/10)] / 60 = 0.19 (19%). Note that if you had not contacted some of the non-respondents, you would have estimated the proportion answering 'yes' as 4/35 = 0.11 (11%). Calculate the variance of the proportion as if you hadn't contacted the non-respondents. Using equation 11, the variance is therefore [(200-35)/200]×(0.11×0.89/34) = 0.0024. From section 5.3, a 95% confidence interval for the proportion experiencing offensive odours is then 0.19 ± (2×0.049) = [0.09, 0.29] - that is, 9 to 29%. This confidence interval will be a bit wider than it should be - it's possible to calculate a variance that takes account of the contacted non-respondents but it's a little more complicated and with an indicative survey, there's really no point in being that accurate when calculating the variance.*[34]

# 7. Council: definitive survey

## *7.1 Sampling frame*

A list of ratepayers (households and businesses) in the area of interest will be the best sampling frame. You could still take a simple random sample (or systematic sample) from this list, and then simply follow the guidelines in section 6 above. You would probably take a larger sample for a definitive survey, to increase the precision of estimates for legal purposes. [Indicative surveys typically use small samples and as a result, 95% confidence intervals are likely to be rather uninformative - see section 5.3]. Remember that if a definitive survey of individuals is necessary and an over-estimate of the severity of an odour problem will help your cause, then you should consult a professional. You may need to randomly sample individuals within households and businesses (see section 4).

In a definitive survey you may wish to interview people face-to-face rather than phoning, so you can ask more complex questions (perhaps on the FIDOL factors).[35] In this section, I will assume you have decided to visit households and businesses, interviewing 'the person most often at home' from each selected household and a suitable representative from each selected business (see section 4). If a selected address turns out to be a flat, then you will end up interviewing a tenant rather than the owner of the property. The tenant 'most often at home' will usually be the most appropriate person to answer your questions.

Often you will want to make estimates for different parts of a population - perhaps for both those inside and outside an affected area. The easiest way to do this efficiently is by using stratified random sampling (section 7.2). You divide the population up into parts (strata), so that later you can make estimates for each stratum. Ideally each stratum is as similar within and as different between strata as possible (that is, similar and different in terms of the community's perception of odour).

In practice, the result might be something like this: you form say three strata as concentric rings around a point source (assuming local topography doesn't affect odour or its influence is unknown). The closest stratum is the area in which most of those who have complained live. The second stratum out has the odd complainant. And the third stratum is an area outside your population of interest, so you can show you've covered everyone likely to find odour a problem (see section 3). You might divide up some or all of these circles into sectors, making more strata, based on your knowledge of the prevailing winds. Do complaints seem more prevalent 'down-wind' of a suspected odour source? Remember, the trick is to form another stratum only if a those living in a particular area are likely to have a very different perception of the problem.

## *7.2 Selecting the sample*

It doesn't pay to have too many strata, because you now have to list all the ratepayers in each stratum. You then take a simple random sample (or systematic sample) of ratepayers in each stratum, using the methods given in section 5.2. Use a different random number to start each systematic sample.

The big advantage of stratified sampling is that it can be far more efficient than simple random sampling (that is, a smaller sample size can give estimates of the same precision). To calculate the overall sample size, first identify the most important question or questions in your survey. For each question, you need to consider the width of the confidence interval you want to end up with. For each question, you need to guess the population variance in each

stratum, using the methods in section 6.2. Once you have this information, you can calculate the overall sample size. Having done this for those questions you consider important, you take the largest overall sample size as the required sample size and allocate this sample among the strata.

Calculate the overall sample size for a given question as:[36]

$$n = \frac{\left(\sum_{h=1}^{L} N_h \sigma_h\right)^2}{\left(N^2 \partial^2 / Z^2\right) + \sum_{h=1}^{L} N_h \sigma_h^2},$$

$n$ = overall sample size;

$N$ = overall population size; 13.

$N_h$ = population size in stratum h (h = 1, ..., L);

$Z$ = 2.0 for a 95% confidence interval;

$\partial$ = half the width of the desired confidence interval;

$\sigma_h^2$ = population variance in stratum h;

$\sigma_h$ = square root of $\sigma_h^2$.

Now allocate this sample size among the strata:[37]

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h}, \qquad 14.$$

$n_h$ = sample size allocated to stratum h.

Note that sometimes $n_h$ turns out to be larger than $N_h$. If this happens, you'll need to make a slight adjustment to this method.[38]

This method of allocation means that if you are taking systematic samples, each stratum will have a different value of k (the integer part of $N_h/n_h$). There are other methods of allocating the overall sample size which keep k constant, but the method shown here (known as Neyman allocation) ensures that strata where responses are more varied are sampled more intensively. This improves sampling efficiency.[37]

Simple random sampling within each stratum ensures that you can make estimates (of averages or proportions) for any single stratum, as well as overall estimates for all those surveyed. However, if you want to find a confidence interval for a single stratum, as a rule of thumb that stratum needs a minimum sample size of 30 (section 5.3).

*7.2.1 You are going to ask respondents if they have experienced any offensive odours in the last week. You want a 95% confidence interval (with width plus or minus 5%) for the proportion that answer 'yes' to this question. You form three strata - concentric rings around the factory you believe is releasing odours. The first column in the table below gives your conservative guess of the proportion likely to answer 'yes' to your question in each stratum. You can then estimate the variance in each stratum as $P_h.(1-P_h)$ - see equation 6. The second column shows the population in each stratum - the total population is 600. Using equation 13, the total sample size is $(231.7^2)/(225+94) = 168$. The third column in the table shows how equation 14 allocates the total sample size among the three strata. Note that the sampling fraction $(n_h/N_h)$ is highest in the stratum where you think more people are finding odour offensive. Note too that the sample size in each stratum is over 30, so you will be able*

*to calculate confidence intervals for each of your strata. You will want to increase the sample size in each stratum to allow for non-response (see section 6.2.2).*

| *Stratum* | $P_h$ | $\sigma_h^2$ | $N_h$ | $n_h$ |
|---|---|---|---|---|
| *Inner ring* | *0.5* | *0.25* | *100* | *36* |
| *Middle ring* | *0.3* | *0.21* | *200* | *67* |
| *Outer ring* | *0.1* | *0.09* | *300* | *65* |
| *Total* | | | *600* | *168* |

### *7.3 Estimation*

Once the survey has been completed, estimate an average as:[39]

$$\hat{\overline{Y}} = \overline{y}_{st} = \frac{\sum_{h=1}^{L} N_h \overline{y}_h}{N},$$

$\hat{\overline{Y}}$ = estimate of population average; 15.

$\overline{y}_{st}$ = average of a stratified random sample;

$\overline{y}_h$ = sample average (equation 7) in stratum h.

The variance of this average is:[40]

$$v(\overline{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h}{n_h} \cdot (N_h - n_h) \cdot s_h^2,$$

$v(\overline{y}_{st})$ = variance of $\overline{y}_{st}$; 16.

$s_h^2$ = sample variance in stratum h.

Use the statistical function of a scientific calculator to find the sample variance in each stratum ($s_h^2$), or calculate this by hand using equation 9 in section 6.3.

Estimate a proportion as:[41]

$$p_{st} = \frac{\sum_{h=1}^{L} N_h p_h}{N},$$

$p_{st}$ = proportion for a stratified random sample; 17.

$p_h$ = sample proportion (equation 10) in stratum h.

The variance of this proportion is:[42]

$$v(p_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 (N_h - n_h)}{(N_h - 1)} \cdot \frac{p_h (1 - p_h)}{(n_h - 1)},$$ 18.

$v(p_{st})$ = variance of $p_{st}$.

When using equations 7, 9 or 10 - to calculate the sample average, variance or proportion in stratum h - just use sample observations from stratum h. That is:

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}, \quad s_h^2 = \frac{\sum_{i=1}^{n_h} y_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} y_{hi}\right)^2}{n_h}}{n_h - 1}, \quad p_h = \frac{a_h}{n_h}. \qquad 19.$$

*7.3.1 You carry out the survey described in section 7.2.1. In the inner ring, 12 out of 30 respondents answer 'yes', they have experienced offensive odours in the last week. In the middle ring 15 out of 60 respondents say 'yes', and in the outer ring 5 out of 60 respondents say 'yes'. Your estimate of the overall proportion is (40+50+25)/600 = 0.19. The variance of this proportion is $(58.5+89.4+93.5)/600^2 = 6.706 \times 10^{-4}$. So a 95% confidence interval for the proportion is 0.19 ± (2.0×0.026) = [0.14, 0.24]. You could also calculate a 95% confidence interval (section 6.3) for say the inner ring as 0.40 ± (2.0×0.076) = [0.25, 0.55].*

### *7.4 Non-response*

To allow for those you cannot contact, use the procedure outlined in section 6.4 to adjust your estimate of the average. Take a random sample of those who were not available in your first attempts to get an interview. You could do this in just one stratum or take a random sample in each of several strata. Make a real effort to interview these non-respondents. Adjust the estimate of the stratum average using the procedure in section 6.4. Use adjusted stratum averages instead of the normal stratum averages when combining estimates from each stratum to give the average of a stratified random sample (in section 7.3). This works for proportions too.

You can use a similar adjustment to account for those who refuse to be interviewed. This method is called the 'basic question' approach.[43] You should identify one question (maybe two questions at most) that best summarises what your survey is about. Because this question needs to be asked to both respondents and non-respondents in as similar circumstances as possible, ideally this 'basic question' is one of the first you ask in your questionnaire. This approach will not be possible if the most important question is 'buried' in the questionnaire to de-sensitise the issue.

When someone refuses to participate, you ask them to answer just this one 'basic question'. Later adjust each stratum average (or proportion) and use adjusted stratum averages instead of the normal stratum averages in your calculation of the average for a stratified random sample.

The adjustment to be made to each stratum average is:

$$\bar{y}_h' = \frac{\left(n_{1h}\bar{y}_{1h} + n_{2h}\bar{y}_{2h}\right)}{n_h},$$

$\bar{y}_h'$ = adjusted sample average for stratum h;
$n_{1h}$ = those who complete a full interview in stratum h;
$\bar{y}_{1h}$ = average answer to the basic question for those in stratum h who complete a full interview; 20.
$n_{2h}$ = those who refuse a full interview in stratum h;
$\bar{y}_{2h}$ = average answer to the basic question for those in stratum h who answer just this one question.

7.4.1 *For the survey described in section 7.2.1, your sample size calculations suggest you need a sample of 36 from the inner ring. You increase this to a sample of 50 to allow for a 70% response. But when you carry out the survey, you manage to complete 20 full interviews in the inner ring. Of the remaining 30, 15 were not at home the <u>three</u> times you called, 5 refused to talk to you and 10 refused an interview but gave answers to just the 'basic question'. This question was whether they had experienced any offensive odours over the last week. In the full interviews, 15 out of 20 said 'yes', but only 2 of the 10 answering just this 'basic question' said 'yes'. Your adjusted estimate of the proportion answering 'yes' in the inner ring is then [20×(15/20) + 15×(2/10)] / (20+15) = 0.51. If you had just used information from full interviews, you would have estimated this proportion as 15/20 = 0.75. You now need to repeat this process for the other two strata, and then you can use these adjusted proportions in equation 17. You can use the data from both sources when calculating the variance (equation 18) - in the inner ring, you have data for your 'basic question' from 20 + 10 = 30 respondents in total.*

For both overall and single stratum estimates, compare the adjusted average for this 'basic question' with an average calculated the usual way. The comparison will indicate whether non-response is likely to cause appreciable bias in estimates for other questions. If appreciable bias seems likely, you can get a statistician to adjust estimates for other questions. [Estimates for other questions can be adjusted by regression, using the information you have just collected about your 'basic question'.[43]]

7.4.2 *For the survey described in section 7.2.1, you require estimates for each ring (as well as overall estimates). Section 7.4.1 shows that in the inner ring, there's a large difference between estimates for the 'basic question' with and without the extra information (0.51 versus 0.75). This difference suggests appreciable bias is possible in estimates for other questions. You should get a statistician to adjust the estimates for these other questions. Otherwise you could be seriously over-stating the problem.*

# 8. Odour producer: indicative survey

## *8.1 Sampling frame*

As an odour producer, you will not have a list of ratepayers from which you can draw a sample. A telephone survey will be quicker and cheaper than household interviewing. Remember to keep your questionnaire short and simple (section 6.1). If you are lucky, your population of interest will coincide with a local calling area. You can simply take a systematic sample from a phone book. More commonly, the population of interest will be only a small part of a local calling area, or spread over several local calling areas. In this case talk to Telecom Directories, Directory Information Services. They can select all the phone numbers within a certain geographic area. They may not be able to exactly match your area of interest - you should ask for the smallest area that completely includes your area of interest. They will sell you a list of phone numbers, without names and addresses (because of the Privacy Act 1993), and at the time of writing their prices seemed reasonable.[44] The rest of section 8 covers this situation.

One other alternative is to contract a council to run the survey for you, using a questionnaire that is acceptable to both you and the council. [The council is unlike to be able to give you a sample of their ratepayers, because of the Privacy Act 1993].

## *8.2 Selecting the sample*

Your population of interest is contained within a local calling area in a phone book, or within a list provided by Telecom's Directory Information Services. The problem is some of the numbers in your sampling frame are for people living outside the area you are interested in. You need to take a sample of telephone numbers, and just telephone those in your population of interest. But even if you have an address as well as a phone number, you may not know if a particular household is in your population of interest. In this case, when you phone you have to find this out first, before you ask the rest of your questions. You only want to record answers for those who are in your population of interest. This situation is called sampling a subpopulation or domain.[45]

To calculate the required sample size, identify the most important question or questions in your survey. For each question, consider the width of the confidence interval you want to end up with. Calculate an approximate sample size for each question as:

$$n_0 = \frac{Z^2\sigma^2}{\partial^2},$$

$n_0$ = first approximation to required sample size;

$Z$ = 2.0 for a 95% confidence interval; 21.

= 1.3 for an 80% confidence interval;

$\partial$ = half the width of the desired confidence interval;

$\sigma^2$ = population variance (as yet unknown).

Of course you don't know the population variance before surveying, and the trick is to make an intelligent guess. Some ideas on how to do this are in section 6.2.

Now you need to adjust this approximate sample size given the number in your population of interest ($N_j$). You probably won't know what this number is: work out the number of phone

numbers in your sampling frame (N), and roughly estimate the proportion of these that belong to households in your population of interest ($p_j$). Multiply these two numbers together as an estimate of $N_j$. You will be conservative if you make $N_j$ slightly larger than you think it probably is.

$$n_j = \frac{n_0 N_j}{n_0 + N_j},$$

$n_j$ = required sample size; 22.

$N_j$ = size of population of interest,

$= N \cdot p_j$.

Remember, $n_j$ should be at least 30 so that later you can calculate reliable confidence intervals (see section 5.3).

You now know how many households you want to sample from the population of interest. But to find these households, you need to take a sample from the sampling frame and then reject those households that turn out not to be in your population of interest. So the sample you initially select (n) needs to be larger than the sample you want to end up with ($n_j$):

$$n = \frac{n_j}{p_j}. \quad 23.$$

*8.2.1 You think around 10% of the population will find a particular odour offensive. You want a 95% confidence interval with width plus or minus 10%. To be slightly conservative (section 6.2.2), you assume that the population proportion (P) is 0.15. Your initial sample size is $2.0^2 \times 0.15 \times 0.85 / 0.1^2 = 51$. There are 300 phone numbers in your sampling frame, and you expect 80% of these numbers to belong to those in your area of interest. So the size of the population of interest ($N_j$) is $300 \times 0.80 = 240$. The required sample size is then adjusted to $51 \times 240 / (51+240) = 42$. But to find these 42 people, you will need to select $42/0.8 = 53$ phone numbers from the sampling frame - say 75 phone numbers to allow for a 70% response.*

You need to follow through this process for each important question and take the largest value of n as the sample size required for your survey.

Now take a simple random or systematic sample of size n from the N phone numbers in your sampling frame (section 5.2). When you phone each number in the sample, first check that the household you've called is in your population of interest. You'll never get exactly the sample size you planned ($n_j$), but you should get close enough. But if your population of interest makes up only a small percentage of the sampling frame, you'll need to phone a lot of numbers to get the (roughly) $n_j$ you need.

### 8.3 Estimation

You can treat this sample as a simple random sample. Use the equations in section 6.3 to calculate averages, proportions and their variances. Provided you just collect responses from those in your population of interest, these equations work if you replace n with $n_j$, and N with $N_j$.[45] That is, use the sample size and population size of your population of interest, not the sampling frame sample size and population size. Now that you have selected the sample, you

are in a position to make a better estimate of $N_j$:

$$N_j = \frac{n_j}{n} \cdot N. \qquad 24.$$

*8.3.1 Of the 75 phone numbers in section 8.2.1, you make contact with 60 households, and 45 of these turn out to be in your population of interest. In 10 out of the 45 households, 'the person most often at home' had noticed an offensive odour during the last week. Your estimate of $N_j$ is therefore (45/60)×300 = 225 - your estimate before taking the sample was 240. Your estimate of the proportion is 10/45 = 0.22 (equation 10) with variance [(225-45)/225]×0.22×0.78/44 = 0.0031 (equation 11). An 80% confidence interval (equation 2) is 0.22 ± (1.3×0.06) = [0.15, 0.29].*

### *8.4 Non-response*

You may wish to sample some of those you are initially unable to contact. By finding some of these people, you can adjust estimates to reduce the bias that results from non-response. The method is a variation on equation 12 and it's not that easy. At this point you either consult your friendly statistician, or try to make some sense out of the following example.

*8.4.1 Of the 75 phone numbers in section 8.2.1, you manage to make contact with only 40 households, of which 30 turn out to be in your population of interest. Of these 30, 6 have noticed an offensive odour during the last week. From the remaining 35 phone numbers, you take a simple random sample (or systematic sample) of 15 and pursue these at length. Of these 15, 10 households turn out to be in your population of interest, and of these 10, 3 have noticed an offensive odour. Your estimate of the proportion (equation 12) is [30×(6/30) + $n_2$×(3/10)] / [30 + $n_2$]. The trouble is, you don't know $n_2$ - the number in your population of interest among those you haven't contacted. Your best estimate of $n_2$ is 35×(10/15). This makes your estimate of the proportion (6 + 7) / 53.3 = 0.24. If you hadn't sampled non-respondents, you would have estimated this proportion as 6/30 = 0.20. Calculate the variance of the proportion as if you hadn't sampled non-respondents. The resulting confidence interval will be a bit wider than it needs to be - it's possible to calculate a variance that takes account of the non-respondents you finally contacted but things are complicated enough already.*[34]

# 9. Odour producer: definitive survey

## *9.1 Sampling frame*

You could use the indicative sampling frame (section 8.1) for a definitive survey. You would probably take a larger sample for a definitive survey, to increase the precision of estimates for legal purposes. [Indicative surveys typically use small samples and as a result, 95% confidence intervals are likely to be rather uninformative - see section 5.3].

In a definitive survey you may wish to interview people face-to-face rather than phoning, so you can ask more complex questions (perhaps on the FIDOL factors).[35] In this section, I will assume you have decided to visit households and businesses, interviewing 'the person most often at home' from each selected household and a suitable representative from each selected business (see section 4).

There are a number of ways you can construct a sampling frame with addresses: no one method will suit all circumstances. The size of the area you want to survey will largely determine which of the following methods is most suitable. In order from small scale to large, possible sampling frames are: (1) a list of houses and businesses from aerial photographs; (2) a list of houses and businesses for each streets in a town or city suburb; (3) the phone book for a local calling area; (4) the 'meshblock' classification system used by Statistics New Zealand.

The first method is very small scale. Using aerial photos, simply number the houses in the area of interest. Take a simple random sample of these numbers using a computer.

The second method requires a list of all the addresses of households and businesses for every street in your area of interest. It will be easiest to assemble this sampling frame in a computer spreadsheet. You can get all the streets (for your population of interest) off a map, and then visit each street to find the last number in the street. This will tell you roughly how many houses there are in the street. Or if this is too much work, visit a Registrar of Electors (New Zealand Post) and look at the 'Habitation Index' for the electorate. The 'Habitation Index' has streets in alphabetical order, and the names and addresses of registered voters who live in each street.[46] Find each street of interest in the 'Habitation Index', and record the last address in the street. Obviously, the 'Habitation Index' will not be very useful for a new subdivision. Once you know roughly the last address in the street, you know roughly how many households are in the street. With a spreadsheet, you can then generate a number for every household in every street and this is your sampling frame.

This trick of using the last address to tell you how many households are in a street is not exact - but often you'll be close enough. In a mathematical sense, the number of the last address in the street will under-estimate the number of households in the street. If the 'Habitation Index' does not contain a high percentage of all the addresses in the street, use the following unbiased estimate:[47]

$$\hat{N}_s = \left( \frac{(m+1)Y_{max}}{m} \right) - 1, \qquad 25.$$

$\hat{N}_s$ = an estimate of the number of houses in a particular street;

$Y_{max}$ = the largest number among the m numbers listed in the Habitation Index for that street.

In a practical sense, even this alternative (equation 25) is likely to under-estimate the number of households in the street - because multiple households at the same street number will be more frequent than parks and empty sections. So your sampling frame will under-represent those living in flats and retirement villages and if these households represent a significant part of your population, you'll just have to visit those streets where there are a lot of multiple households at the same street number and record all the letter box numbers in these streets. You can get a pretty good idea of whether a street has a high proportion of flats or a retirement village from the names and addresses in the 'Habitation Index'.[48]

A sampling frame of this sort will contain some street addresses that simply do not exist, and will miss out some addresses that do exist. The best you can do is to ensure that these additions and subtractions are as 'random' as possible. That's why you should visit and amend your sampling frame with actual letter box numbers where multiple households are prevalent or in newly subdivided areas. Remember to document your procedures so that later you can explain what you did and why. You should also 'post-stratify' estimates (see section 9.4). If your sampling frame under-represents certain groups, estimates may be biased - with sensible 'post-stratification', adjusted estimates will be less biased.

The third method uses the phone book. Since the phone book gives addresses, you could use this for a sampling frame. But it's more likely your population of interest does not lie within a single local calling area; or your population of interest is only a small part of a local calling area; or you want to make estimates for different parts of your population of interest. In each case, the phone book is not going to be a satisfactory sampling frame. [Telecom's Directory Information Services can only supply you with phone numbers, not addresses.]

The fourth method is surveying on a large scale. Statistics New Zealand uses a 'meshblock' classification system to divide the country up into areas roughly the size of a city block (rural meshblocks tend to be larger). In theory, you could take a sample of meshblocks from those containing your population of interest, then list all the households in each of the sampled meshblocks, and sample some of these households and businesses. This is called cluster sampling. Several stages of sampling are involved: first a sample of meshblocks, then a sample of houses and businesses within each sampled meshblock. While this is the best way to survey a large or high density area (such a city), you are going to need a statistician. Calculating the required sample size, calculating averages, proportions and their variances are all more difficult with a cluster sample.

One other alternative is to contract a council to run the survey for you, using a questionnaire that is acceptable to both you and the council. [The council is unlike to be able to give you a sample of their ratepayers, because of the Privacy Act 1993].

### *9.2 Selecting the sample*

With numbered houses on an aerial photo, use a computer to draw a simple random sample. Do not use a systematic sample in this situation because you are unlikely to number the houses in a random order (see section 5.2). On the other hand, take a systematic sample if you are using the phone book as the sampling frame. Use section 6.2 to calculate the sample size for a simple random or systematic sample.

If you construct a sampling frame of street names and house numbers in a spreadsheet, you could take a simple random sample, or use a systematic sample provided the streets are in alphabetical order. Or you could arrange the streets into strata, and take a simple random sample (or systematic sample) from each stratum. Each stratum should be as similar within and as different between as possible (that is, similar and different in terms of the community's

perception of odour). Read section 7.1 on how to divide a population up into strata; read section 7.2 on how to calculate the sample size for a stratified random sample.

### 9.3 Estimation

Use section 6.3 to calculate estimates from a simple random or systematic sample. Use section 7.3 to calculate estimates from a stratified random sample.

### 9.4 Non-response

To account for those you cannot contact, use the method described in section 6.4 (for simple random sampling) and in section 7.4 (for stratified random sampling). A random sample of those not available in first attempts at an interview is used to adjust the estimate of an average or proportion.

You can use a similar adjustment to account for those who refuse to be interviewed. This 'basic question' approach is described in section 7.4. You should identify one question (maybe two questions at most) that best summarises what your survey is about. Because this question must be asked of both respondents and non-respondents in as similar circumstances as possible, this 'basic question' has to be one of the first you ask in your questionnaire. When someone refuses to participate, ask if they will just answer this one question. Answers from those who complete a full interview and from those who answer just the 'basic question' are then combined to give an adjusted estimate of an average or proportion.

The next method of adjustment can reduce not only the bias due to non-response (section 5.4), but also the bias in estimates due to inadequacies in a sampling frame. So this method of adjustment will be particularly useful if you've had to construct the sampling frame yourself, using a street map and the 'Habitation Index' (see section 9.1).

This third method is called 'post-stratification'. But you need to think about its use before you survey, because you have to find out which 'post-strata' each respondent belongs to by asking the appropriate questions in your survey. 'Post-strata' typically involve groups based on say age, sex or ethnicity. Like the usual sort of geographically based strata (section 7.1), 'post-strata' should be as similar within the group and as different between groups as possible - similar and different with respect to perceptions of odour. While you could form 'post-strata' within geographical strata, this would involve a large overall sample size. Each 'post-stratum' needs a sample size of at least 20 [49], so you are most likely to use 'post-stratification' in conjunction with simple random sampling. There is a more efficient way to 'post-stratify' across (rather than within) geographical strata, but the calculations are not for the faint-hearted.[50] I will just consider 'post-stratification' as it applies to simple random sampling.

You need to know the frequency with which each group occurs in your population of interest. The easiest way to find this out is using Supermap - a Statistics New Zealand database on CD ROM. You will find Supermap at major public libraries, polytechnics and universities. Using Supermap, you can identify the meshblocks (see section 9.1) that make up your population of interest. Supermap will give the number of people in the meshblock at the last Census, by age, sex, ethnicity and many other variables.

So to use this method, perceptions of odour should vary between groups and you have to be able to get data from Supermap for each of these groups. Groups based on age or on work status (full time, part time, or not in the labour force) are likely to fit these two criteria. [Even if perceptions don't vary between groups, 'post-stratification' won't increase the bias in estimates.] If those in flats and retirement villages make up a significant part of your

population and you think they are likely to be under-represented in your sampling frame (and this may bias estimates), then form groups based on age. If you are more concerned about bias in estimates because of low response rates, then form groups based on work status. You might form groups based on other variables - it depends on what sort of people you think are under-represented in your sample (either because of problems with your sampling frame or because of non-response). Since you have to ask questions in your survey to establish group membership, it makes sense to look at the way these question were asked in the last Census. You may also need to check the definitions used in the last Census - for concepts such as 'part time' or 'not in the labour force'.[51]

To adjust estimates, replace the usual sample average ($\bar{y}$ in equation 7) with:[52]

$$\bar{y}_w = \sum_{g=1}^{G} \frac{N_g}{N} \cdot \bar{y}_g,$$

$\bar{y}_w$ = post - stratified estimate of the sample mean;

$N_g$ = number of people in group g; 26.

N = sum up number of people in all G groups;

$\bar{y}_g$ = sample average for those in group g.

Since a proportion is just a special sort of average, you can replace the averages in the above equation with the appropriate proportions.

The 'post-stratified' estimate weights each group average by the frequency with which that group occurs ($N_g/N$). That's why it doesn't matter too much if meshblock boundaries don't coincide exactly with the boundaries of your population of interest. If there's a slight mismatch, it probably won't change the group frequencies much. 'Post-stratification' reduces both the bias in estimates, and the variance. So you can use the usual variance calculation (equation 8 or 11) because the result will be conservative. You could get a statistician to calculate a more accurate variance or to 'post-stratify' using several variables (perhaps using both age and work status) or to 'post-stratify' across geographical strata.[43,50]

9.4.1 *You are going to ask respondents if they have experienced any offensive odours in the last week. You identify all the streets and part streets in your population of interest from a street map. You use the 'Habitation Index' and visits to construct a sampling frame, and you then take a simple random sample of 80 out of 400 households. You think your sampling may under-represent flats in this lower socio-economic area. From Supermap, you find that the area you're interested in includes most of six meshblocks. You add up the people in these six meshblocks, for each of three age groups: 0-29, 30-59, 60 and over. In your questionnaire, you ask respondents which age group they belong to and after the survey, you estimate a sample proportion for each age group ($p_g$). The data are in the table below. The 'post-stratified' estimate for the proportion who say 'yes' is (0.50×0.33)+(0.30×0.24)+(0.20×0.18) = 0.27. The usual sample proportion is 20 / 80 = 0.25.*

| *'Post-stratum'* | $N_g$ | $N_g/N$ | *'yes'* | $n$ | $p_g$ |
|---|---|---|---|---|---|
| *0-29* | *600* | *0.50* | *7* | *21* | *0.33* |
| *30-59* | *360* | *0.30* | *10* | *42* | *0.24* |
| *60+* | *240* | *0.20* | *3* | *17* | *0.18* |
| *Total* | *1200* | *1.00* | *20* | *80* | |

# 10. Notes

[1] Department of Statistics. (1992) *A guide to good survey design.* Wellington: Department of Statistics.

[2] Isaaks, EH; Srivastava, RM. (1989) *An introduction to applied geostatistics.* Oxford: Oxford University Press. On p 108: 'In a study of the concentration of some pollutant, for example, we are not really interested in the average concentration of the pollutant in the samples we have collected. What we actually want to know is the concentration of the pollutant over some larger region.'

[3] Ministry for the Environment (1994) *Discussion document: odour measurement and management.* Wellington: Ministry for the Environment, p14.

[4] *Auckland Regional Authority v Mutual Rental Cars.* (1987) 2 NZLR, p647-681.

[5] Ministry for the Environment (1995) *Odour management under the Resource Management Act.* Wellington: Ministry for the Environment, p54.

[6] *Ibid.*, p11.

[7] Conditions under which survey results would be judged admissible have included 'That the proper universe was examined' or 'interviewees must be selected so as to represent a relevant cross-section of the public' - *Auckland Regional Authority v Mutual Rental Cars.* (1987) 2 NZLR, p680.

[8] Sarndal, C-E; Swensson, B; Wretman, J. (1993) *Model assisted survey sampling.* New York: Springer-Verlag, p137.

[9] Kish, L. (1965) *Survey sampling.* New York: John Wiley, p398.

[10] Fowler, FJ. (1985) *Survey research methods.* Beverly Hills: Sage Publications, p33.

[11] Asking for the household member with the nearest birthday gives too much control to the person first contacted. This person often claims to be the person with the nearest birthday when they are not; this person may not know the birthdays of all the others in the house; or they may decide that visitors are not part of the household. Or the interviewer may decide that it is easier to interview this first contact rather than someone else who isn't perhaps at home. Other methods of selecting a household member force the interviewer to do a proper job.

[12] Sarndal, C-E; Swensson, B; Wretman, J. (1993) *Model assisted survey sampling.* New York: Springer-Verlag, p139-140.

[13] Kish, L. (1965) *Survey sampling.* New York: John Wiley, p400, 403.

[14] Sarndal, C-E; Swensson, B; Wretman, J. (1993) *Model assisted survey sampling.* New York: Springer-Verlag, p151.

[15] *Ibid.*, p73.

[16] Cochran, WG. (1977) *Sampling techniques (third edition).* New York: John Wiley, p212-213.

[17] *Ibid.*, p217-219.

[18] Cochran, WG. (1977) *Sampling techniques (third edition).* New York: John Wiley, p27.

[19] Berenson, ML; Levine, DM; Rindskopf, D. (1988) *Applied statistics: a first course.* Englewood Cliffs, New Jersey: Prentice Hall, p227.

[20] Department of Statistics. (1992) *A guide to good survey design.* Wellington: Department of Statistics, p33.

[21] Fowler, FJ. (1985) *Survey research methods.* Beverly Hills: Sage Publications, p147.

[22] *Ibid.*, p68-69.

[23] *Ibid.*, p66-67.

[24] Berenson, ML; Levine, DM; Rindskopf, D. (1988) *Applied statistics: a first course.* Englewood Cliffs, New Jersey: Prentice Hall, p259-261.

[25] *Ibid.*, p266-268.

[26] Cochran, WG. (1977) *Sampling techniques (third edition).* New York: John Wiley, p76.

[27] Equation 5 follows if what you are trying to measure is distributed according to a normal distribution. Then 99.9% of observations (ie. almost all) lie within 3 standard deviations on either side to the mean. So the range of what you observe is about 6 standard deviations - and the variance is just the square of a standard deviation. If you think that what you are trying to measure should follow a rectangular or triangular distribution, use similar methods - see Cochran, WG. (1977) Sampling techniques (third edition). New York: John Wiley, p81.

[28] For a variable taking only values zero or one (see section 5.3), the population variance is $N/(N-1)\times P\times(1-P)$. Since $N/(N-1)$ will be close to one, the population variance is therefore approximately $P\times(1-P)$ - see Cochran, WG. (1977) Sampling techniques (third edition). New York: John Wiley, p51.

[29] Cochran, WG. (1977) *Sampling techniques (third edition).* New York: John Wiley, p20-21.

[30] *Ibid.*, p26.

[31] *Ibid.*, p51.

[32] *Ibid.*, p52.

[33] *Ibid.*, p57-60.

[34] *Ibid.*, p370-371

[35] Ministry for the Environment (1995) *Odour management under the Resource Management Act.* Wellington: Ministry for the Environment, p14.

[36] Cochran, WG. (1977) *Sampling techniques (third edition).* New York: John Wiley, p98.

[37] *Ibid.*, p 98.

[38] *Ibid.*, p104.

[39] *Ibid.*, p91.

[40] *Ibid.*, p95.

[41] *Ibid.*, p107.

[42] *Ibid.*, p108.

[43] Bethlehem, JG; Kersten, HMP. (1985) On the treatment of nonresponse in sample surveys. *Journal of Official Statistics* 1(3), 287-300.

[44] On 1/11/96 the cost (GST exclusive) of selecting phone numbers was 3.5 cents per number with a set up cost of $150. Alternatively, Directory Information Services would take a random sample for you, for 18.5 cents per number and with no set up cost. With both these alternatives you had to buy a minimum of 3000 numbers. So if you need only a small sample for an indicative survey, these prices aren't really relevant - phone 0800-501-515 for price and product options. Some time in the future, you may be able to specify your area of interest using Statistic New Zealand's meshblock classification system (see section 9.1).

[45] Cochran, WG. (1977) *Sampling techniques (third edition).* New York: John Wiley, p34.

[46] The 'Habitation Index' itself is not suitable as a sampling frame - too many eligible voters do not register, and so many households are missing. But from the 'Habitation Index' you can get an idea of how many households are in each street, and using this information you can put together your own sampling frame.

[47] Rayner, JCW. (1994) Estimating Saddam's arsenal. New Zealand Statistician, 29(2) 59-65.

[48] You could also use Supermap (section 9.4) to find areas with a high proportion of flats or retirement villages. 'Dwelling type' data include percentages for 'flats or houses joined together' and for 'homes for the elderly'.

[49] Sarndal, C-E; Swensson, B; Wretman, J. (1993) *Model assisted survey sampling.* New York: Springer-Verlag, p267.

[50] *Ibid.*, p268-269.

[51] Department of Statistics. (1991) *1991 Census of Population and Dwellings: Concepts, Definitions and Classifications.* Wellington: Department of Statistics.

[52] Cochran, WG. (1977) *Sampling techniques (third edition).* New York: John Wiley, p134.

# 11. Appendix

In section 4, I suggested that if a definitive survey of individuals is necessary and an over-estimate is going to be to your advantage, then you will need to sample individuals within households and businesses. I now want to summarise how I think it would be easiest to survey the community in a way that does not over-estimate individual concern. The material in this appendix is more difficult than the material in the rest of this document - you may find you need to consult a statistician. A statistician will typically use more efficient (but even more complex) methods than those I'm going to describe here. Professional advice is recommended because it can be cost-effective. You pay for a statistician, but you may end up taking a smaller sample to get the same accuracy. But following the great do-it-yourself tradition, I will now expand on the brief comments I made at the end of section 4. As in section 7, I will assume you have decided to interview face-to-face, and that you are using stratified random sampling.

*Briefly, the common method of selecting the individual who has the next birthday is not recommended.*[1] *Kish gives a better method: the interviewer ranks household members by increasing age (the interviewer doesn't actually need to know these ages), then uses a random number table to choose one person from this list.*[2]

The best way to do this is to attach a sticky label to the top of each questionnaire. The label should look something like this:

| **Paper** | **Number of individuals in the household** | | | | | |
|---|---|---|---|---|---|---|
| **Number** | **1** | **2** | **3** | **4** | **5** | **6** |
| | **Order individuals by decreasing age, and select the...** | | | | | |
| **137** | **1** | **1** | **1** | **4** | **3** | **5** |

The bottom line of the sticky label has a selection of random numbers. If the household has five individuals in it, the interviewer would interview the third eldest person in the household. The interviewer should circle the number of individuals in the household. This is important - it serves as both a record of how many individuals are in the dwelling, and which individual was selected (perhaps this individual is not available, and another interviewer will have to call later). If there are more than six individuals in the household, the interviewer should note down the number of individuals in the household, and then just use the selection indicated by the column for six individuals.

A table at the end of this appendix has enough random numbers for 100 sticky labels. If you need more, just cycle through the table again. The label above uses the random numbers corresponding to Paper 37 in the table.

*Estimates need to be adjusted to take account of the variable selection probabilities.*

Equations 15 and 17 give estimates of averages and proportions for a stratified random sample. With a second stage of selection (where an individual is chosen at random from within each selected household or business) these equations need to be modified.

Equation 15 combines sample averages from each stratum - the $\bar{y}_h$. Each stratum average is just the usual average:

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h},$$

$y_{hi}$ = each sample observation in stratum h; 27.

$n_h$ = the number sampled in stratum h.

With a second stage of selection, use the following 'ratio estimate' (as it's called) to estimate the average in each stratum. If the number of individuals in a dwelling is m, then:[3]

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} m_{hi} y_{hi}}{\sum_{i=1}^{n_h} m_{hi}},$$

$y_{hi}$ = the sampled observation from household i, in stratum h; 28.

$m_{hi}$ = the number of individuals in household i, in stratum h.

Equation 16 combines sample proportions from each stratum. With a second stage of selection, estimate each stratum proportion as the ratio:

$$p_h = \frac{\sum_{i=1}^{n_h} m^*_{hi}}{\sum_{i=1}^{n_h} m_{hi}},$$

$m^*_{hi}$ = the number of individuals from household i, in stratum h
**if the observation from that household** 29.
**has the value one rather than zero - see section 5.3**;

$m_{hi}$ = the number of individuals in household i, in stratum h.

*11.1.1 You sample 20 households out of the 50 in a particular stratum. [Note that the sample is too small to calculate reliable confidence intervals for this particular stratum - see section 7.2.] In each of the 20 households, a person is chosen at random. Five of these 20 people have experienced offensive odours in the last week. These five live in households with 1, 2, 2, 3 and 5 individuals respectively. In total, there are 80 individuals living in the 20 households. Using equation 29, the proportion of individuals who have experienced offensive odours in this stratum is (1+2+2+3+5)/80 = 0.16 - not the usual estimate of 5/20 = 0.25 (equation 10).*

*The variance in these estimates can be calculated as if there were no second stage of sampling. This will under-estimate the true variance,*[4] *but Kish suggests the difference will not be large provided most households and businesses are about the same size (say one to six people).*[5] *An appreciable number of larger businesses in the population of interest means that businesses should be treated differently. A sample of people will need to be selected from each business, and second stage variances calculated for the business component of the community. Alternatively, businesses could be sampled with replacement (see section 5.2) and then there is no need to calculate second stage variances.*[6]

The easiest way to estimate the variance will be to separate out large businesses. In one stratum, put the households and small businesses (where say 6 or fewer are employed); in another stratum put large businesses. You could make this split in just one or perhaps several of your geographically based strata.

In 'large business' strata, sample businesses with replacement. If you draw the same business twice, you will want to select two individuals from that business. Use simple random sampling or systematic sampling (section 5.2) to select individuals from within a large business. Try to make sure that the answers of the first person interviewed do not influence the answers of the second person interviewed.[2]

If you use this approach, you can then calculate variances (equations 16 and 18) as if there were no second stage selection. In 'household and small business' strata, the under-estimate of the variance will be small; in 'large business' strata, with replacement sampling ensures that there is no need to calculate second stage variances.

[1] Asking for the household member with the nearest birthday gives too much control to the person first contacted. This person often claims to be the person with the nearest birthday when they are not; this person may not know the birthdays of all the others in the house; or they may decide that visitors are not part of the household. Or the interviewer may decide that it is easier to interview this first contact rather than someone else who isn't perhaps at home. Other methods of selecting a household member force the interviewer to do a proper job.

[2] Kish, L. (1965) *Survey sampling.* New York: John Wiley, p398

[3] *Ibid.*, p400.

[4] Sarndal, C-E; Swensson, B; Wretman, J. (1993) *Model assisted survey sampling.* New York: Springer-Verlag, p139-140.

[5] Kish, L. (1965) *Survey sampling.* New York: John Wiley, p400, 403.

[6] Sarndal, C-E; Swensson, B; Wretman, J. (1993) *Model assisted survey sampling.* New York: Springer-Verlag, p151.

## Table for selecting an individual at random from households and small businesses

| Paper Number | Number of adults in household | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 1 | 2 | 2 | 3 | 5 | 5 |
| 2 | 1 | 1 | 2 | 4 | 3 | 2 |
| 3 | 1 | 1 | 1 | 1 | 3 | 4 |
| 4 | 1 | 2 | 1 | 2 | 2 | 1 |
| 5 | 1 | 2 | 3 | 4 | 4 | 2 |
| 6 | 1 | 2 | 3 | 3 | 4 | 3 |
| 7 | 1 | 2 | 2 | 4 | 3 | 4 |
| 8 | 1 | 2 | 3 | 1 | 5 | 4 |
| 9 | 1 | 1 | 3 | 3 | 2 | 6 |
| 10 | 1 | 1 | 2 | 1 | 4 | 4 |
| 11 | 1 | 2 | 1 | 4 | 2 | 6 |
| 12 | 1 | 2 | 1 | 2 | 1 | 6 |
| 13 | 1 | 2 | 1 | 1 | 5 | 2 |
| 14 | 1 | 2 | 2 | 2 | 5 | 2 |
| 15 | 1 | 1 | 3 | 3 | 1 | 5 |
| 16 | 1 | 2 | 1 | 2 | 5 | 1 |
| 17 | 1 | 1 | 1 | 3 | 1 | 6 |
| 18 | 1 | 1 | 3 | 1 | 5 | 4 |
| 19 | 1 | 2 | 2 | 2 | 1 | 6 |
| 20 | 1 | 2 | 2 | 2 | 4 | 6 |
| 21 | 1 | 2 | 2 | 2 | 3 | 6 |
| 22 | 1 | 1 | 1 | 4 | 5 | 4 |
| 23 | 1 | 1 | 2 | 4 | 3 | 1 |
| 24 | 1 | 2 | 3 | 3 | 4 | 2 |
| 25 | 1 | 2 | 3 | 4 | 5 | 4 |
| 26 | 1 | 2 | 2 | 1 | 3 | 3 |
| 27 | 1 | 1 | 2 | 4 | 2 | 6 |
| 28 | 1 | 1 | 3 | 2 | 1 | 4 |
| 29 | 1 | 1 | 1 | 2 | 4 | 3 |
| 30 | 1 | 2 | 3 | 3 | 3 | 6 |
| 31 | 1 | 2 | 3 | 1 | 4 | 3 |
| 32 | 1 | 2 | 3 | 3 | 4 | 5 |
| 33 | 1 | 1 | 2 | 1 | 5 | 4 |
| 34 | 1 | 1 | 2 | 1 | 2 | 2 |
| 35 | 1 | 2 | 1 | 3 | 2 | 4 |
| 36 | 1 | 1 | 2 | 1 | 4 | 1 |
| 37 | 1 | 1 | 1 | 4 | 3 | 5 |
| 38 | 1 | 2 | 1 | 4 | 4 | 1 |
| 39 | 1 | 1 | 3 | 2 | 3 | 5 |
| 40 | 1 | 1 | 1 | 4 | 2 | 5 |
| 41 | 1 | 1 | 1 | 2 | 1 | 4 |
| 42 | 1 | 1 | 3 | 4 | 2 | 3 |
| 43 | 1 | 2 | 3 | 2 | 1 | 2 |
| 44 | 1 | 2 | 1 | 2 | 4 | 6 |
| 45 | 1 | 1 | 2 | 1 | 4 | 5 |
| 46 | 1 | 2 | 1 | 1 | 2 | 5 |
| 47 | 1 | 1 | 3 | 3 | 5 | 4 |
| 48 | 1 | 1 | 2 | 1 | 1 | 6 |
| 49 | 1 | 2 | 2 | 3 | 2 | 5 |
| 50 | 1 | 2 | 1 | 4 | 3 | 1 |

| Paper Number | Number of adults in household | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 51 | 1 | 2 | 3 | 3 | 1 | 6 |
| 52 | 1 | 2 | 2 | 3 | 1 | 1 |
| 53 | 1 | 2 | 3 | 3 | 4 | 4 |
| 54 | 1 | 1 | 1 | 2 | 1 | 2 |
| 55 | 1 | 2 | 2 | 1 | 5 | 6 |
| 56 | 1 | 1 | 2 | 3 | 3 | 1 |
| 57 | 1 | 2 | 2 | 2 | 3 | 3 |
| 58 | 1 | 1 | 2 | 3 | 1 | 6 |
| 59 | 1 | 2 | 3 | 2 | 2 | 1 |
| 60 | 1 | 2 | 3 | 2 | 3 | 2 |
| 61 | 1 | 2 | 3 | 3 | 5 | 5 |
| 62 | 1 | 1 | 3 | 2 | 5 | 3 |
| 63 | 1 | 2 | 3 | 2 | 2 | 6 |
| 64 | 1 | 2 | 3 | 2 | 3 | 5 |
| 65 | 1 | 1 | 1 | 3 | 2 | 2 |
| 66 | 1 | 1 | 3 | 2 | 1 | 6 |
| 67 | 1 | 1 | 3 | 1 | 2 | 4 |
| 68 | 1 | 2 | 1 | 1 | 2 | 5 |
| 69 | 1 | 2 | 1 | 4 | 4 | 3 |
| 70 | 1 | 2 | 3 | 3 | 5 | 6 |
| 71 | 1 | 1 | 1 | 3 | 4 | 6 |
| 72 | 1 | 2 | 3 | 2 | 1 | 1 |
| 73 | 1 | 2 | 3 | 1 | 5 | 5 |
| 74 | 1 | 1 | 2 | 3 | 5 | 3 |
| 75 | 1 | 1 | 1 | 4 | 4 | 2 |
| 76 | 1 | 2 | 3 | 3 | 4 | 3 |
| 77 | 1 | 1 | 3 | 3 | 1 | 4 |
| 78 | 1 | 2 | 2 | 2 | 4 | 5 |
| 79 | 1 | 2 | 3 | 3 | 1 | 4 |
| 80 | 1 | 1 | 2 | 4 | 3 | 1 |
| 81 | 1 | 1 | 3 | 2 | 5 | 4 |
| 82 | 1 | 2 | 1 | 4 | 5 | 3 |
| 83 | 1 | 2 | 3 | 4 | 5 | 4 |
| 84 | 1 | 2 | 3 | 3 | 3 | 6 |
| 85 | 1 | 1 | 1 | 4 | 2 | 6 |
| 86 | 1 | 1 | 3 | 3 | 5 | 6 |
| 87 | 1 | 2 | 3 | 3 | 4 | 2 |
| 88 | 1 | 2 | 3 | 3 | 3 | 5 |
| 89 | 1 | 1 | 2 | 2 | 1 | 3 |
| 90 | 1 | 1 | 3 | 4 | 1 | 4 |
| 91 | 1 | 1 | 2 | 1 | 4 | 4 |
| 92 | 1 | 1 | 2 | 1 | 3 | 3 |
| 93 | 1 | 1 | 3 | 1 | 5 | 5 |
| 94 | 1 | 2 | 1 | 1 | 4 | 5 |
| 95 | 1 | 1 | 2 | 2 | 1 | 1 |
| 96 | 1 | 1 | 1 | 1 | 4 | 4 |
| 97 | 1 | 2 | 2 | 2 | 4 | 3 |
| 98 | 1 | 1 | 1 | 3 | 2 | 3 |
| 99 | 1 | 1 | 1 | 1 | 3 | 4 |
| 100 | 1 | 1 | 3 | 3 | 3 | 4 |