

Estimation and Probabilistic Linkage in Sample Surveys of Anonymous Organisations

A thesis
submitted in fulfillment
of the requirements for the degree of
Master of Science in Statistics
by
Nicholas Jury
Supervisor:
Richard Arnold
2017

School of Mathematics and Statistics
Victoria University of Wellington
P.O. Box 600
Wellington
New Zealand

ABSTRACT

Drug use takes on many forms, normally this will be just the occasional alcoholic drink, certain individuals drug use develops into habitual use, or more extreme drugs, and then into full addiction. Some of these addicted individuals realise the harmful nature of their addiction and join the anonymous support group, Narcotics Anonymous.

This study focus' on the creation of population size estimates, and an estimate of the size of the persistent population between two survey years. These estimates are created from the 2004 and 2008 surveys run by the Narcotics Anonymous Fellowship, as this is an anonymous organisation with no register of the membership database maintained.

Population size estimation for an anonymous organisation is established using simulation methods. The bootstrap estimation was used to estimate characteristics about the two populations. Probabilistic matching was used to identify individuals who were in both the 2004, and 2008 surveys. Once identified, a logistic regression model was used to establish what impacts an individual to remain in the programme.

Factors that impacted an individual being persistent in the population included the individual education, employment status, and if they had worked through all the 12 steps of Narcotics Anonymous.

ACKNOWLEDGMENTS

I would like to thank Dr Richard Arnold for his patience and support throughout the study. Along with Sharleen Forbes who assisted as great co-supervisor.

I have also appreciated the support from the crew at Aspeq, who have supported me with the balance between work and study.

I have also appreciated the Hutt Valley and Wellington taekwondo community who have always encouraged me on. I will also like to thank my family for their support.

I would like to acknowledge this study would not be possible without assistance from Narcotics Anonymous, and the permission of the New Zealand Narcotics Anonymous Regional Service Committee for their permission to use the data from their two surveys.

CONTENTS

1. <i>Introduction</i>	7
1.1 Background on Narcotics Anonymous	7
1.2 Objectives of the Study	9
1.2.1 Aim of Study	9
1.2.2 Review of Previous Study	11
1.2.3 Structure of Thesis	11
2. <i>Data Source</i>	13
2.1 Target Population	13
2.2 Survey Design	13
2.2.1 Questionnaire Design	15
2.2.2 Potential Errors	17
2.3 Description of variables	20
2.3.1 Demographic Information	20
2.3.2 Recovery Information	20
2.3.3 Education and Employment Information	22
2.3.4 Health Information	23
2.4 Overview of Data Results	24
3. <i>Cross-Sectional Estimation</i>	37
3.1 Cross-Sectional Estimation Methodology	37
3.1.1 Sample Design	37
3.1.2 Weights	39
3.1.3 Point Estimates	40
3.1.4 Confidence Intervals	41
3.1.5 Improved Estimates	43
3.1.6 Improved Confidence Interval	45
3.2 Results	46
3.2.1 2004 Population Size Estimate	46
3.2.2 2008 Population Size Estimate	47
3.3 Bootstrap Estimation	49
3.3.1 Bootstrap	49

3.3.2	Bootstrap Methodology	50
3.3.3	Point Estimates	50
3.3.4	Confidence Intervals	51
3.3.5	Bootstrap Estimation	52
3.4	Statistical Theory	55
3.4.1	Weighted Point Estimates	55
3.4.2	Regression	55
3.5	Other Cross-Sectional Estimates	56
3.5.1	Linear Regression	56
3.6	Concluding Remarks	61
4.	<i>Record Linkage Theory</i>	62
4.1	Deterministic Matching	63
4.2	Probabilistic Matching	65
4.2.1	Slack	65
4.2.2	u -probability	66
4.2.3	m -probability	68
4.2.4	Weight	68
4.2.5	Selecting Matching Variables	72
4.2.6	Quality of Matches	75
5.	<i>Matching The Two NA Surveys</i>	77
5.1	Grouping the variables	78
5.2	Clean Time Prior to 1995	81
5.2.1	Determining u -probability	81
5.2.2	Estimating m -probability	84
5.2.3	Determining Slack	86
5.2.4	Determining m -probability	91
5.2.5	Selecting Match Variables	95
5.3	Clean Time Prior to 2000	102
5.4	Complete Dataset	106
6.	<i>Longitudinal Analysis</i>	110
6.1	Logistic Regression	110
6.2	Model Selection	110
6.2.1	Criterion-based procedure	110
6.2.2	Testing-based procedure	111
6.3	Logistic Regression Analysis	112
6.4	Longitudinal Analysis	116
6.4.1	Persistent population estimate	119

6.4.2	Standard errors of the estimate of the persisting population	121
6.4.3	Longitudinal Estimates	123
6.4.4	Concluding Remarks	125
7.	<i>Analysis of Matched Data</i>	126
7.1	Probabilistic Matching	126
7.2	Longitudinal Estimates	129
8.	<i>Summary</i>	136
8.1	Results	136
8.2	Further Research	137
8.3	Final Remarks	138
	<i>Appendix</i>	144
A.	<i>Question and Survey Forms</i>	145
B.	<i>Additional Graphs</i>	155

1. INTRODUCTION

Drug use takes on many forms. These range from the occasional alcoholic drink to taking class A drugs such as cocaine. Many people don't struggle with drug use, however, there are those that develop habitual use which grows into a full addiction. This research focuses on the effectiveness of the NA recovery programme in New Zealand, by estimating the number of individuals staying in the programme and what factors contribute to an individual staying in the programme. This is done by analysing the results of the 2004 and 2008 Narcotics Anonymous membership surveys.

1.1 Background on Narcotics Anonymous

Narcotics Anonymous, commonly referred to NA or "The Fellowship", is a global organisation of recovering drug and alcohol addicts. NA's member base consists of both men and women of all ages for whom drug use/abuse has become a major problem in their lives. These individuals are commonly referred to as addicts who have admitted their lives have become dominated by narcotics and their life is controlled by their drug use. This addiction is commonly referred to as a disease by the NA organisation. This disease is out of the member's control but the choice of recovery comes down to the member.

Similar to other anonymous organisations Narcotics Anonymous is one in which the members commit to complete abstinence from all drugs. The members meet regularly in a support system to stay "clean" (drug free) and assist each other in the recovery from the impact on their lives from their addiction.

Like Alcoholics Anonymous, Narcotics Anonymous uses a Twelve Step programme. The member is asked to complete the following steps in order to assist in overcoming their addiction:

1. Admission to being powerless to the addiction, lives had become unmanageable;
2. Believe in higher power to restore sanity;
3. Turn will and lives over to the care of God as member understands Him;
4. Members have to make a searching and fearless moral inventory of themselves;
5. Admit to God, themselves, and to others the exact nature of their wrongs;
6. They were entirely ready to have God remove defects of character;
7. Ask God to remove these shortcomings;
8. Make a list of persons harmed by addiction, and be willing to make amends to them all;
9. Make direct amends to such people whenever possible, except when in doing so would bring injury to them or others;
10. Continue to make personal inventory and when wrong promptly admit it;
11. Sought through prayer and meditation to improve their conscious contact with God, as they understood Him, praying only for knowledge of His will for them and carry that out;
12. Having had a spiritual awakening as a result of these steps, we tried to carry this message to addicts, and to practice these principles in all their affairs.

(Prayer 1976)

A key point to this Twelve Step programme is that it asks the member to make a personal decision for finding and believing in a “power greater than oneself”. Even though the me asks for its members to make this decision, the organisation is not associated with any particular religion, nor is it associated with any other organisation or institution.

Narcotics Anonymous operates by holding meetings, which is a support group for individuals going through the recovery process. These meetings are held weekly with members of Narcotics Anonymous attending regularly in most cases one or more meetings per week.

Using anonymity assists the addicts to attend meetings without fear of legal or social implications. There is also an established atmosphere of equality among its member base ensuring that no one personality or circumstance will be considered more important than any other.

An individual who is still using is welcomed at NA meetings, many members become “clean” and are recovering throughout the programme. If a member is “using” the member is asked to refrain from speaking during a meeting but is welcomed to speak with other members before and after the meeting.

1.2 Objectives of the Study

This study uses the data from the 2004 and 2008 surveys conducted by Narcotics Anonymous. It builds on the previous research of, Estimation in Anonymous Organisation by Richard Arnold and Sharleen Forbes (Arnold & Forbes 2005).

The objectives of the study are as follows:

1. Cross-Sectional Estimation

- Create and report separate summary statistics for both 2004 and 2008 including estimates of population size.

2. Longitudinal Estimation

- Use probabilistic matching to match respondents from both 2004 and 2008 survey.
- Create longitudinal estimates including, the number of individuals who remained, left, and joined NA between 2004 and 2008.

1.2.1 Aim of Study

The primary objective of the study is to establish estimates of the population size of Narcotics Anonymous using the 2004 and 2008 membership surveys.

The analysis is broken into two parts, the first is an investigation of the population size of both years using cross sectional estimates. The second section is determining retention rates of the programme by probabilistic matching (also known as Record linkage, or Fuzzy matching) and creating longitudinal estimates of the population size based on key variables.

The first part of the research presents cross sectional estimates, in particular estimates of the population size of active members, and creates estimates to describe what influences an individual staying in the programme and staying clean.

To create population estimates, Richard Arnold's and Sharleen Forbes' cross sectional approach is used. This research focuses on validating the methods for calculating the estimates of active membership in an anonymous organisation.

In the second stage of the research, the main focus will be on creating estimates of the characteristics of Narcotics Anonymous by a probabilistic matching (commonly referred to as probabilistic record linkage), between the two surveys and changes over time. Unlike in cross-sectional estimation, probabilistic matching focuses on using observations from individuals that could possibly exist in both surveys. The main focus of probabilistic matching is to create a matched dataset, consisting of individuals in both surveys.

Ideally this matched dataset would be formed using a unique identifier. However, in this research there is no unique identifier present, and because of this deterministic matching is not possible. Instead probabilistic matching is used, as in probabilistic matching a series of identifiers are used to calculate the most likely candidates that are a match, and from this the aim of creating longitudinal estimations about the population will be achieved.

Probabilistic matching uses a wide range of identifying characteristics about an individual, and creates a probable match between two or more data sets. Early work on record linkage was done by Halbert L. Dunn and formalised by Fellegi & Sunter (1969).

This had a focus on record linkage using a clear characteristic about the individual (typically a unique identifier, student number, IRD number, etc). Probabilistic matching tackles the problem of when no unique identifier exists. In probabilistic matching, a weight is computed for each identifier based on how well it correctly identifies between a match and a non match.

These weights are based on certain probability principles. Once the matches have been made, estimates about the population size, and properties about the population can be made.

The overall aim for the research is to provide estimates to the Narcotics Anonymous Regional Service Committee on certain characteristics that members of the organisation demonstrate i.e. an individual retention time, clean

time etc.

1.2.2 Review of Previous Study

This research builds on Richard Arnold and Sharleen Forbes technical report, “A method for estimating active membership for organisations that do not maintain registers of members” (Arnold & Forbes 2005). The research around this technical report was conducted on the 2004 Narcotics Anonymous data, and the 2004 Alcoholics Anonymous data.

The report sets out a methodology for estimating the population size of organisations that do not retain registers for their membership base. Two main characteristics are investigated, the Association Time (the length of time since their first attendance), and the commitment time (the length of time they have been actively following the principles of the organisation).

The paper establishes estimates for the 2004 populations based on a sample for Narcotics Anonymous and Alcoholics Anonymous. After these estimates have been established, the paper moves to describe estimates for the total population size using imputation methods for missing results.

Once the population estimates for both Narcotics Anonymous and Alcoholics Anonymous have been established, estimates about characteristic of the population are established. This is done by using linear regression, these characteristics focus on the association time (Time spent in NA/AA) and the commitment time (time individual has been “clean”). A full review of the methodology that was used is given in Chapter 3.

1.2.3 Structure of Thesis

In Chapter Two we discuss and describe the data source, the 2004 and 2008 surveys for NA, and review the previous results from these surveys. Along with this discussion the survey design and a summary of both datasets is done.

Chapter Three establishes the cross-sectional estimation, which includes the methodology for estimating the population of an anonymous organisation. This extends the work done by Richard Arnold and Sharleen Forbes.

Chapter Four lays out the theory of probabilistic matching, which then is applied in Chapter Five, where individuals responding in 2004 and 2008 are matched.

The results from the matching are used in Chapter Six which develops the longitudinal estimation. This includes a logistic regression to establish what factors are associated with an individual staying in the programme.

Chapter Seven summarises the research and the findings made by the longitudinal estimation, with final concluding remarks being contained in Chapter 8.

2. DATA SOURCE

2.1 *Target Population*

NA organises itself into weekly meetings, with a specific day of the week, time and location. The Regional Service Committee do monitor a register of currently active meetings, and has contact information for each meeting

The complexity around this survey design is due to the fact that NA does not retain an active register of their members. In lieu of the active membership database, an up to date register of active groups /meetings is maintained. These meetings are used as a means to survey the population.

At the first stage, if the meeting refused to participate in the survey, then as a consequence all individuals attending that meeting were considered to be refusing to participate in the survey.

The surveys were run over a week. This week is referred to as the survey week and is designed to be a “typical” week of NA meetings.

A “typical” survey week was defined as a week that avoided major events. These events included but were not limited by, school holidays, public holidays, and any major NA event. These NA events are special meetings that are held throughout the year. These events include conventions, camps, and NASC (Needs Assessment and Service Coordination) which is a government funded disability support service.

The week of the surveys for 2004 and 2008 were; Friday 12 November 2004 - Thursday 18 November 2004 and Friday 21 November 2008 - Thursday 27 November 2008.

2.2 *Survey Design*

The survey design is broken down into two main parts:

- The Questionnaire Design;
- Potential Errors.

Each part is discussed in detail below.

The way the survey was conducted was that the Narcotics Anonymous Fellowship contacted each meeting and requested they participate in the survey. This was an optional participation for each meeting, if the meeting decided to respond, a set of surveys were posted out to the meeting contact person and the meeting was classified as a responding meeting. The surveys were posted out during a particular week which met the goal of capturing all the meetings, this survey week having to be a week when all meetings were held.

For every responding meeting an individual at each meeting was asked to be the “scribe”. The scribe’s job was to hand the survey forms to everyone at the meeting and collect them back at the end. In addition the scribe had to count the number of individuals in attendance at each meeting, along with the number of individuals who had already completed the survey. For individuals who had already completed the survey earlier that week due to attending another meeting earlier in the week, the scribe didn’t give them a survey form to fill out.

Once all the forms were collected the scribe filled out an additional form noting down all the attendance information. All these forms were posted back to the Narcotics Anonymous fellowship. This information was then captured from the paper forms and entered into the datasets used in this study.

As the attendance count done by the scribe was carried out during meetings, errors in the count can occur. This is due to meetings not being a settled environment, with individuals moving around constantly in the meeting, getting coffee, going to the bathroom etc. This fluid dynamic behaviour of the meeting makes the scribe’s job difficult.

Using these meetings to survey the individuals we can achieve our goal of gaining an estimate of the population size of NA. However with each meeting the survey designed allowed a meeting non-response. This was in addition to the individual volunteering to participate in the survey.

These surveys had a two stage census survey design. In the first stage each of the meetings were contacted, then in the second stage the individuals attending the meeting on a survey day were contacted.

2.2.1 Questionnaire Design

The surveys questionnaire contained 5 main sections:

- About you
 - Clean Date;
 - First NA Date;
 - Age;
 - Closest city or Town;
 - Ethnic group.
- About your recovery
 - Sponsor information;
 - Attendance frequency;
 - Where did the individual get clean;
 - Information about completing the 12 Steps of the NA programme and other counselling support.
- Before you came into recovery;
 - Drug use information;
 - Drug preferences.
- More about you;
 - Employment information before and after joining NA;
 - Education information before and after joining NA.
- Other Information.
 - Health information;
 - Criminal record due to drug use.

The actual questionnaires can be found in the appendix.

Clean date refers to the date the individual declared themselves clean, where First NA date is the date the individual first attended a NA meeting. Individuals can be clean before starting the programme or become clean after, because of this the individuals Clean Date and First NA Date most likely

will not match. The recovery stated in the questionnaires refers to the individuals path of becoming clean of their drug addiction. The survey asked questions about the individuals drug use habits to find if there was a drug that the individual particularly used the most (Drug use information), or a favoured drug (Drug preferences).

The changes to the 2008 survey were kept to a minimum. The changes that were made, were made as a result of confusion in the 2004 survey. The surveys were designed to be consistent with each other as a link between the two years.

The changes that were made included the response categories for ethnicity, in 2004 the responses, Samoan, Cook Island Maori, Tongan, Niuean, and Fijian being grouped all together as Pacific Peoples in the 2008 survey.

In 2008 there were 2 extra questions added to the “About your recovery” section. These were binary yes/no questions about doing service for the NA programme, and doing residential treatment programme for their addiction.

In 2008 for the question about the individual’s drug use time, “How long did you use drugs for?” changed to “How long did you use drugs in total?”. This question was changed from a fixed categorical question with responses, “Under 1 year”, “Between 1 year and 4 years”, “Between 5 years and 9 years”, “Between 10 years and 14 years”, and “15 years and over”, to an open ended response when the individual could state the exact number of years they used drugs.

The final change was regarding contracting/developing a diagnosed health condition. The change that occurred was from 3 choices:

- No;
- Medical - Please Specify;
- Mental Illness - Please Specify.

In 2008 the questions, “BEFORE your clean date, did you contract/develop a diagnosed health condition?”, and “SINCE your clean date, did you contract/develop a diagnosed health condition?”, were limited to 2 sub questions with “none” being an option in each of the mental, and medical choices. This allowance of “none” covered the choice of “no” in the 2004 survey.

2.2.2 Potential Errors

There are three main types of errors that are analysed in this study.

- Selection Bias;
- Data processing errors;
- Sampling errors.

The survey and analysis is designed to minimize these errors as much as possible.

Selection Bias

In selection bias there are 4 main areas which occur in the study:

- Coverage Bias;
- Non-response bias;
- Question bias;
- Compliance bias/ False response.

Coverage bias or survey coverage errors are a result of missing individuals of the sample that we wish to survey, or including individuals that should not be included. Individuals that may have been falsely included were individuals that are not NA members but are observing the meeting. The survey was designed to avoid collecting information on these observers and all possible avenues were taken to avoid these individuals being captured.

Members that were excluded from the survey included

- Members that do not attend meetings;
- Members that were unable to attend meetings during the survey weeks;
- Members that attend only meetings that refused to participate.

Non-response bias arises when the individual is surveyed but chooses not to respond to the survey. Non-responders were classified if they

- did not complete a survey form;
- did not return their survey form;
- returned their form after the survey date.

In addition to survey non-response, there is the chance that the member did not respond to individual items on the questionnaire.

Question bias is where the question is misunderstood by the respondent. There are many reasons that a question can be misunderstood; possible response are the question could be open to interpretation. The language used in the questions is not standardised. The language is set at too high a reading level for the target population. The question is culturally or religious specific making it difficult for all individuals to answer. All questions in the questionnaire were kept as simple as possible, with a record for misunderstood questions in 2004 which were improved in the 2008 survey.

Compliance bias occurs where the respondent responds with what they think is an appropriate response rather than their actual response. This bias comes about as for certain questions a “true response” is looked at as an unfavourable response. Even in anonymous surveys this bias occurs. There is no way of measuring the extent of this error occurring, In this study we assume that the individual gives correct response even to unfavourable questions. e.g. “Do you have criminal convictions as a result of your drug use?”. This is due to the response being kept as confidential.

Data processing errors

Data processing errors occur at the end of survey process before any analysis starts. These errors occur at random and there is no systematic approach to determine where and when these errors occur.

These errors occur when the data entry team inputs the collected survey information. Types of data processing errors include but are not limited to

- Incorrect field response insertion;
- Mis-keyed results for the correct field;
- Partial insertion or no insertion of response.

An example of mis-keying error is inserting the wrong date into commitment time i.e. actual date 12/12/1990 is inserted as 12/12/1992. Another example is of partial insertion or of no insertion as a responses. A partial insertion could be leaving off the year part of a date, or the data processing team could leave the field response blank when the individual responded.

The chance of these errors were minimized by having the data processing team working in pairs for both surveys to double check the response insertion.

Self-validating computerized data capture systems were also used to assist in minimizing the data processing errors.

Sampling errors

Sampling errors occur in sample surveys whenever properties of the sample of a population are used to make inferences about the population. The difference between the sample estimates and the true population estimates are the sampling error. These errors can be calculated using standard statistical techniques.

As sampling errors can occur any estimates that are derived from the survey data will vary from each potential sample we take. The sampling errors are reflected in the 95% confidence intervals established for the estimate.

2.3 Description of variables

2.3.1 Demographic Information

The demographic information of *Age* and *City* were collected as an open ended response in both 2004 and 2008. Age is grouped in two different ways for the analysis. First the age is grouped in 5 year increments e.g. 30-34 years starting at under 20 years old (< 20 years) to over 60 years old (60+ years). City was further grouped into the different regions; Central, Midland, Northern, and Southern. This grouping was code to the variable *Region*.

Other demographic information that was captured is *Sex*, and *Ethnicity*. These were closed ended responses, with sex being a binary response of either male or female. There were multiple responses for ethnicity, these were collapsed into three main groups; New Zealand European/ Pakeha, Maori, and Other (all remaining responses), then collapsed further into two groups Maori and Non-Maori.

The final variable that is of interest that provides some demographic information on the individual is the *Criminal* variable. This measured a binary response (Yes/No) to the question if they had a criminal conviction due to their drug use.

2.3.2 Recovery Information

The key recovery information statistics of the study are *NA Time*, and *Clean Time*. *NA Time* is also referred to as association time. This is the time since the respondent's first NA meeting to the time of the survey. *Clean Time* is the time between the respondents time they became clean and the survey date, and is the time in years the individual has remained clean.

First NA Date has been used to define the variable *NA Time*. *First NA Date* is separated into *First NA Day*, *First NA Month*, and *First NA Year*. Similarly *Clean time* was calculated by using the *Clean Date*. *Clean Date* is also separated into *Clean Day*, *Clean Month*, and *Clean Year*.

Looking at the plots of *Clean Time* vs. *NA Time* for 2004 and *Clean Time* vs. *NA Time* for 2008 in Figure 2.1:

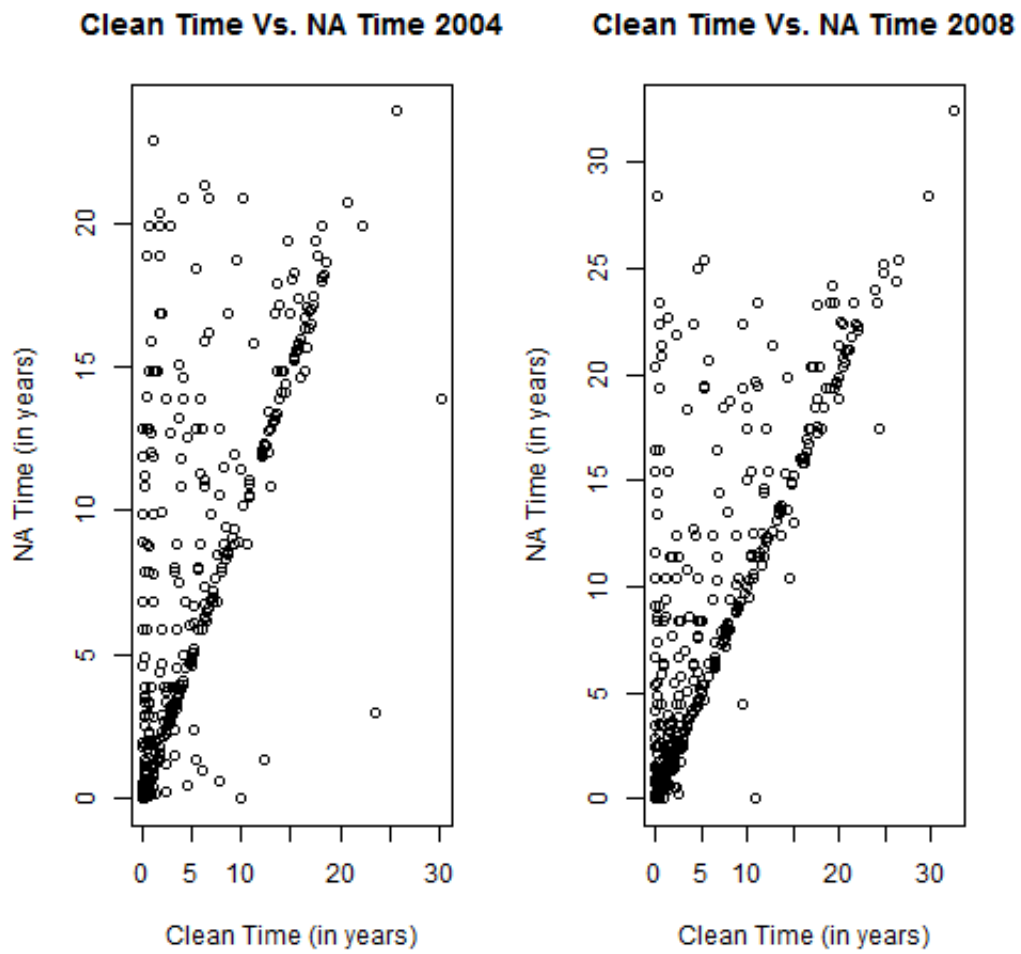


Fig. 2.1: Clean Time vs. NA Time

Individuals with long clean time have been associated with NA for a proportional long time. The converse does hold true; an individual has been associated with NA for a long time does not mean they have a long clean time. This shows that individuals who have been associated with NA for a long time are still susceptible to relapsing.

In NA each member may or may not have a nominated sponsor, who is a support person they can contact whenever they are in need of help in managing their addictions, or to work through the 12 steps.

Sponsor is the coded variable for the individuals response of whether they have a sponsor or not. If an individual has a sponsor the variable *Sponsor Contact* is the measurement of how often they meet up with their sponsor. In addition to having a sponsor the individual can also be a sponsor, this is captured by *Sponsor Others*.

Each respondent in both 2004 and 2008 were asked how often they attended NA meetings. They were given the choice of either putting number per week or per month. This is coded as the *Meeting Frequency* which is calculated as the number of meetings attended per year.

Clean Location was the closed ended response which describes where the respondent became clean.

Each respondent was asked “Have you started to work the 12 steps?” the binary yes/no response was coded as *Start Steps*. *All Steps* is the yes/no response referring to whether the respondent has completed the 12 steps at least once.

Other Fellowships was where the individual stated if they were part of any other 12 step fellowships, for example Alcoholics Anonymous. The variable *Other Support* captures whether the individual is doing other addiction counselling/programmes or support.

Most Influence variable encapsulates the information about what drove the individual to join Narcotics Anonymous in the first place.

2.3.3 Education and Employment Information

Income Source variable describes the employment status of the respondent. The respondents were given a closed ended response ranging from beneficiary to full-time employment these are grouped into 4 main areas Employed, Beneficiary, Student, and Unemployed, and grouped further into Employed or

Unemployed.

Education refers to the highest qualification the respondent holds. This is an ordinal variable ranging from no qualification to having post graduate qualification.

Paid work describes each respondent's type of work. This was a closed ended response which had an open ended other option for the respondent to respond if they didn't fit in any category. Both 2004 and 2008 shared the same categorisations of variables.

All these questions had a counter part which asked the respondent for the same information, but prior to the individual joining Narcotics Anonymous.

2.3.4 Health Information

Two key areas of health are captured, medical condition and mental condition. This information is reflected in the variables, *PreMental*, *PreMedical*, *OngoingMental*, *OngoingMedical*, *PostMental*, and *PostMedical*.

In 2004 the health variables were measured in two parts; the first part measured if the individual has/had a medical or mental condition. The second part then is an open ended response to specify the condition if one was present. The *PreMental*, *PreMedical*, *OngoingMental*, *OngoingMedical*, *PostMental*, and *PostMedical* variables measure the binary yes/no response from the first part. In 2008 the second part is changed from an open ended response to a closed response with the choices of

- None;
- Depression;
- Bipolar;
- Psychosis;
- Other.

for mental illness, and

- None;
- Hep. C;
- Organ Damage;
- HIV;

- Other.

for medical illness.

2.4 Overview of Data Results

An overview of unweighted estimates are used to summarise the initial results from the 2004 and 2008 surveys.

In both surveys there is a near even split between male and female with males taking up 57% of the population of both years, this is illustrated in Figure 2.2. This population split is different to the 2004 NZ population split, which have the near even split with more than half of the population female.

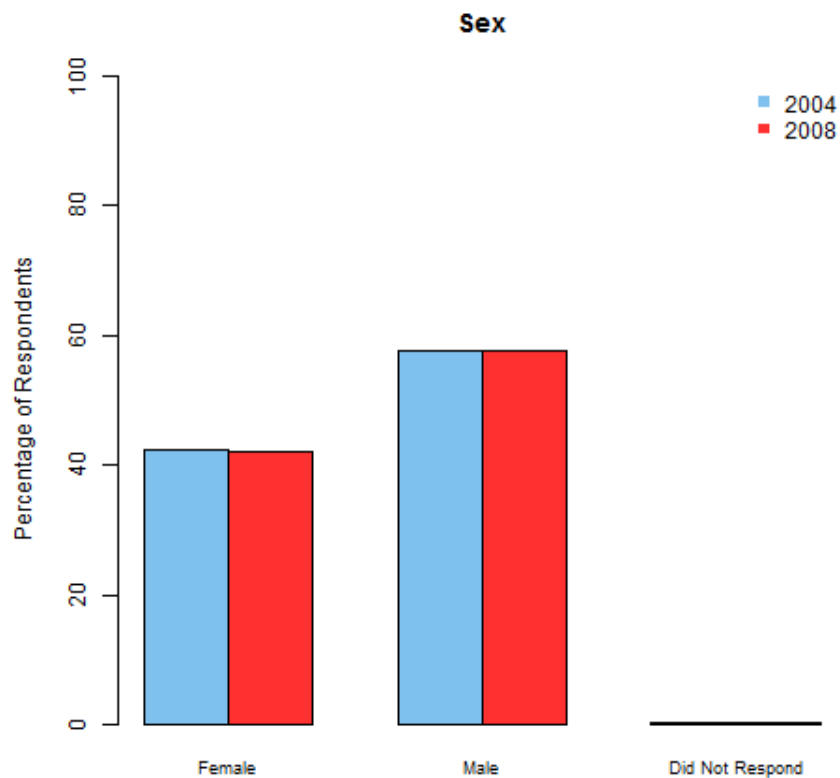


Fig. 2.2: Distribution of the individual's sex (2004/2008)(unweighted)

NZ European/ Pakeha is major ethnicity of the respondents for both 2004 and 2008 (74%, and 75% respectively), illustrated in Figure 2.3. The NA population has a higher proportion of Maori individuals compared to the 2004 NZ population (13.5%)

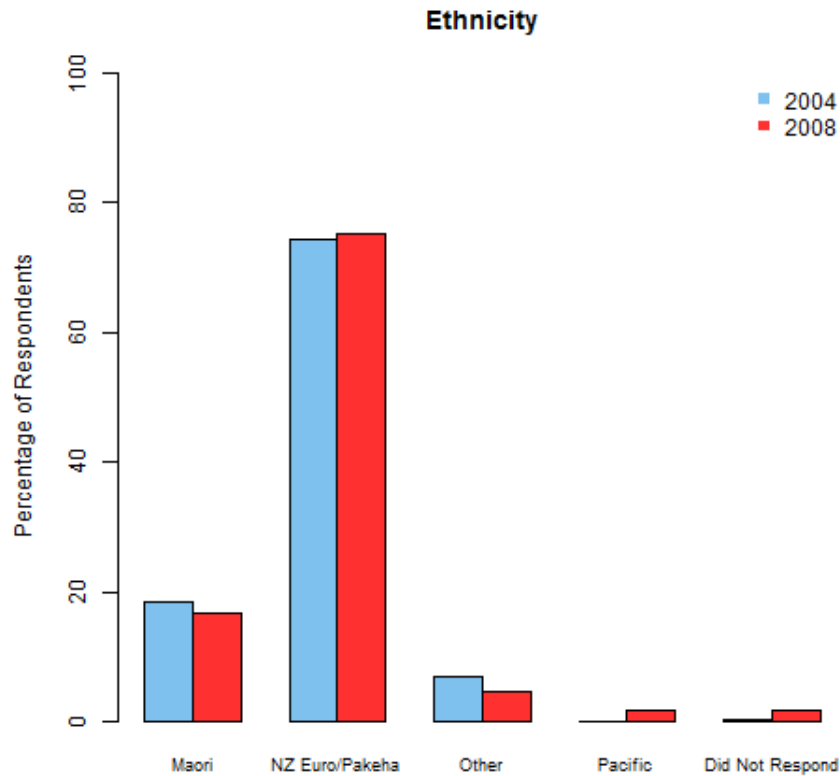


Fig. 2.3: Distribution of an individuals ethnicity (2004/2008)(unweighted)

The majority of respondents (70%) had a sponsor. Roughly half of these individuals meet up with their sponsor at least weekly (see Figure 2.4).

Approximately 70% of 2004 respondents, and approximately 75% of 2008 respondents, also responded that they didn't sponsor other members, shown in Figure 2.5.

In 2004 there was a near even split in respondents who where part of other 12 steps fellowships, in 2008 there was a decrease to near 40% for individuals who where part of other 12 steps fellowships. In both surveys the majority of participants did not attend attentional support.

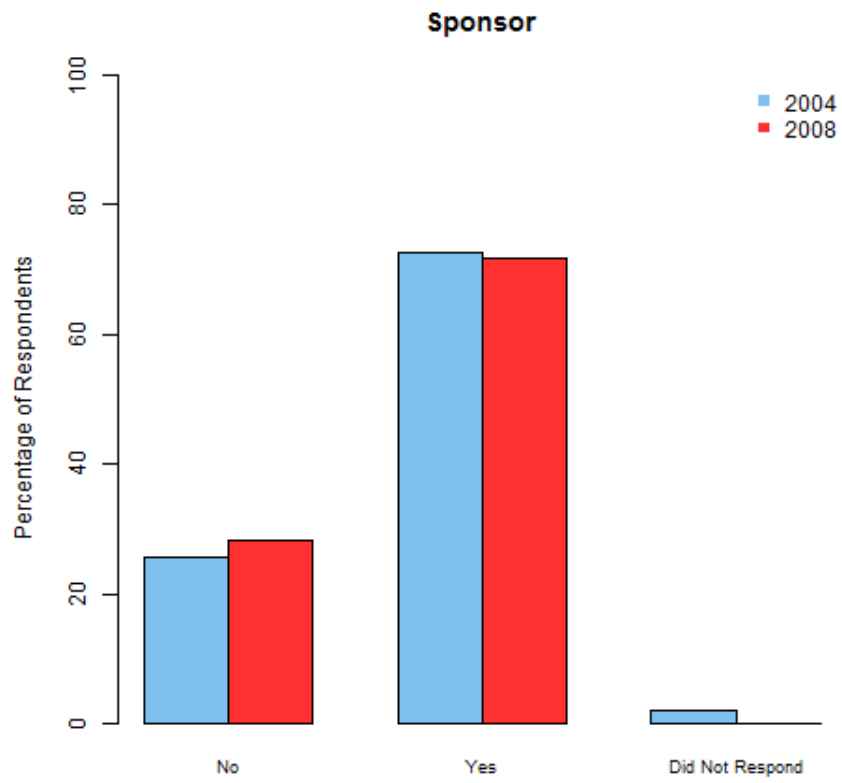


Fig. 2.4: Distribution of individuals who have a sponsor (2004/2008)(unweighted)

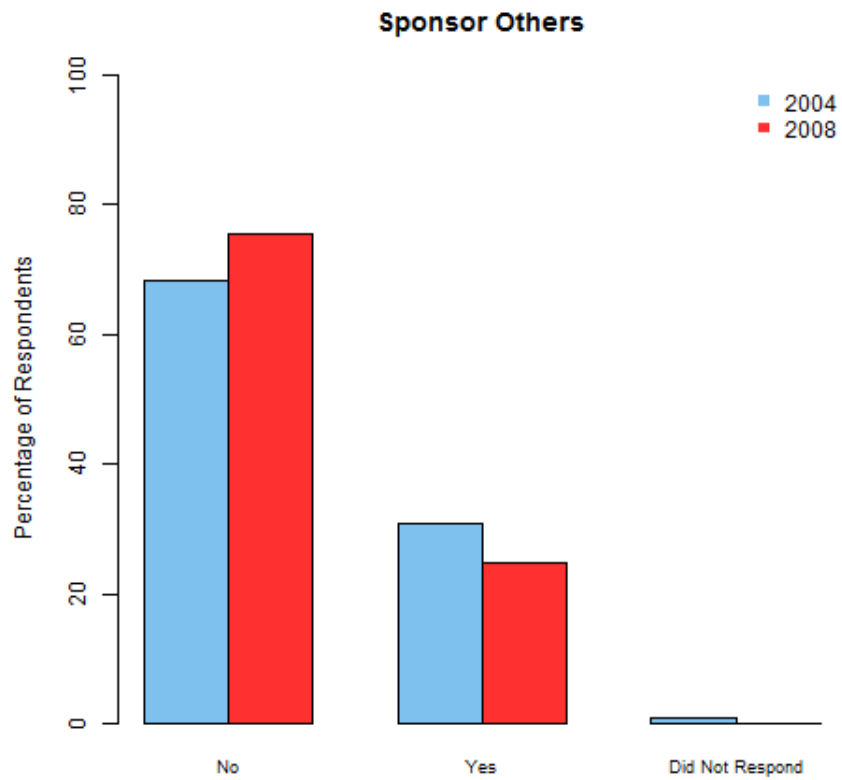


Fig. 2.5: Distribution of individuals who sponsor other members (2004/2008)(un-weighted)

In both years “treatment” was the main location where the individual became clean as shown in Figure 2.6. With most of the individuals stating the most influential people to inspire the individual to join NA, was other NA members or the treatment facility itself.

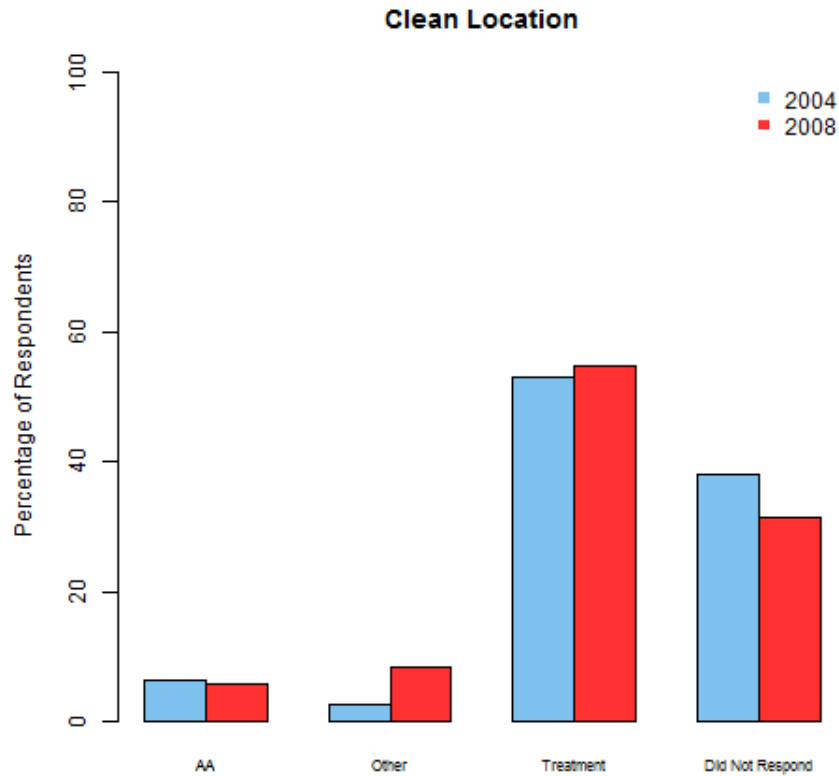


Fig. 2.6: Distribution of where the individual become clean (2004/2008)(un-weighted)

In both 2004 and 2008 at least 87% of respondents had started working their way through the 12 step programme (Figure 2.7) with less than half actually having completed all 12 steps (2.8) by the time the survey was done.

The drug use time for the participants in both surveys was 15 years or more, this was followed by 10-14 years. Individuals who had used drugs for less than 5 years had the lowest attendance during the survey weeks (Figure 2.9).

In 2004 individuals responding did not have a drug of choice, which was replaced by other drug as the drug of choice in 2008. The most common

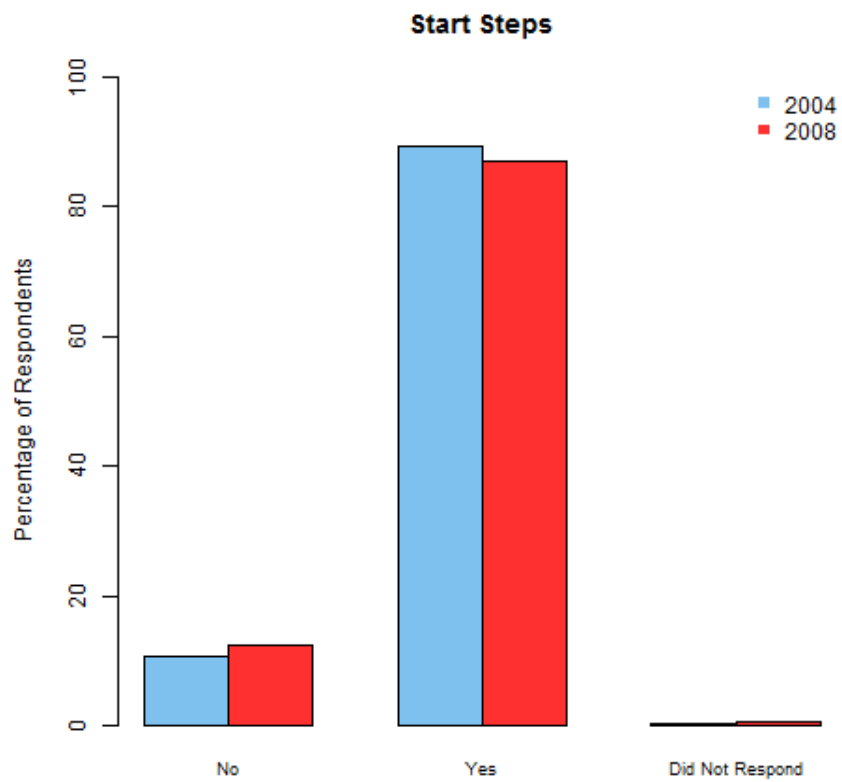


Fig. 2.7: Distribution of individuals who have started the 12 step programme (2004/2008)(unweighted)

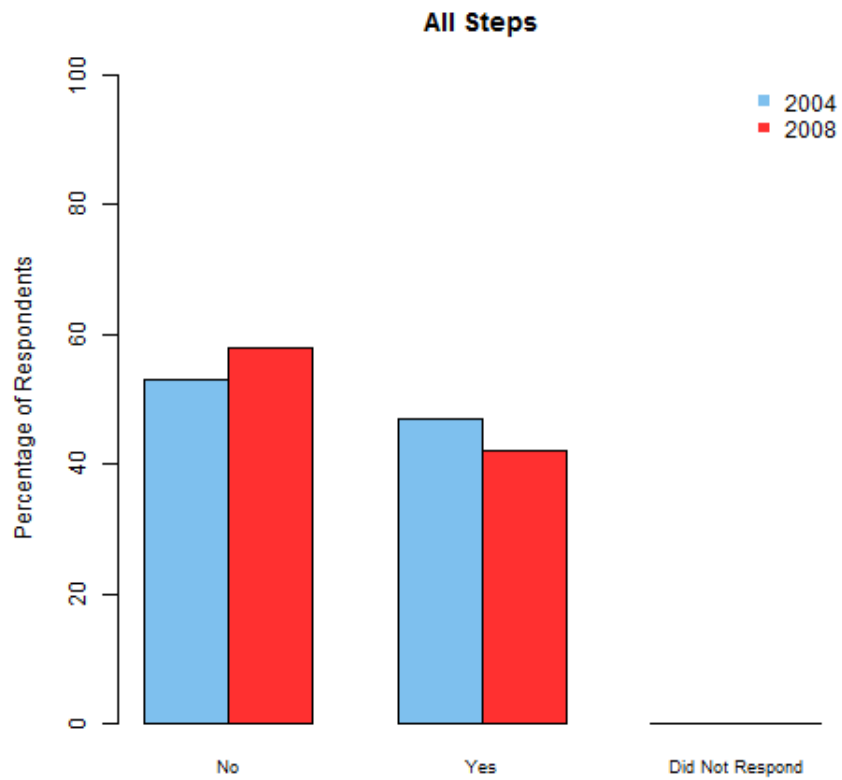


Fig. 2.8: Distribution of individuals who have completed all the 12 steps (2004/2008)(unweighted)

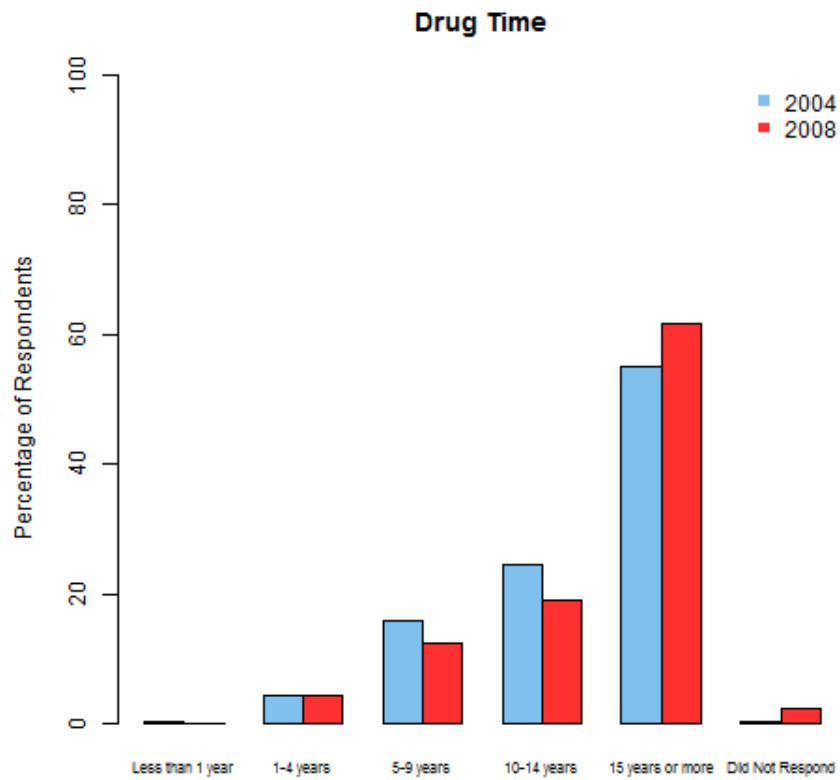


Fig. 2.9: Distribution of the length of time an individual used drugs (2004/2008)(unweighted)

drugs that respondents consumed were cannabis and alcohol in both the surveys.

Employed individuals were the highest attending in both survey years followed by beneficiaries. Even though the highest group was employed at the time of the survey, before joining Narcotics Anonymous most respondents were beneficiaries or already in full time employment.

Prior to joining NA the most common education level was either school or no education, this is shifted to individuals having Tertiary, or Graduate level education after joining NA (Figure 2.10). For 2004 the NA population had a lower proportion of individuals with no education and a higher proportion of individuals with tertiary education.

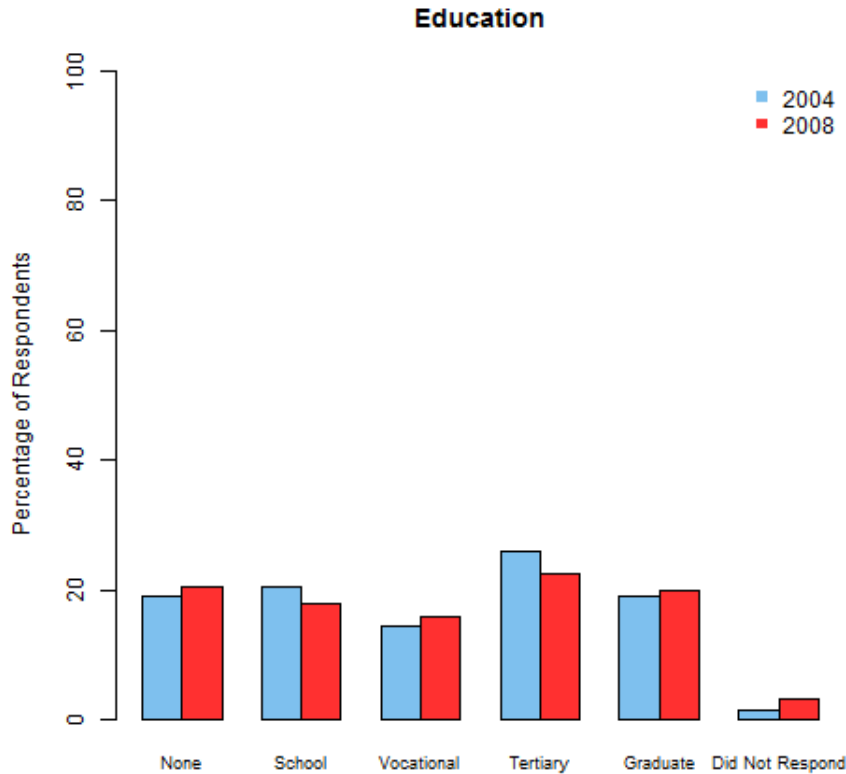


Fig. 2.10: Distribution of individual's level of education (2004/2008)(unweighted)

There is no clear cut profession that is in common for the the respondents either pre or post joining Narcotics Anonymous. The three main professions

which capture the majority of the respondents are:

- None;
- Trade/Industry;
- Service/Clerical.

For 2004 the majority of individuals had no prior mental/medical issues before joining Narcotics Anonymous, in 2008 the majority of individuals said they had prior mental problems but no medical problems (Figure 2.11). At the time of the survey at least 70% of individuals had no ongoing mental/medical issues (Figure 2.12), with 75% of respondents not having any mental/medical issues post joining NA (Figure 2.13).

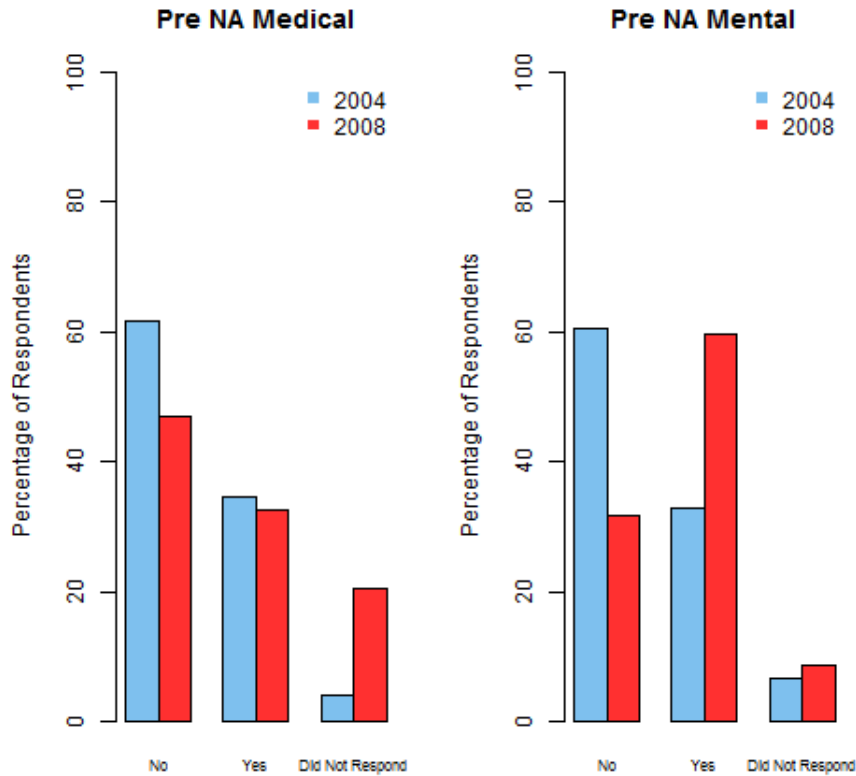


Fig. 2.11: Distribution of individual's health condition before joining NA (2004/2008)(unweighted)

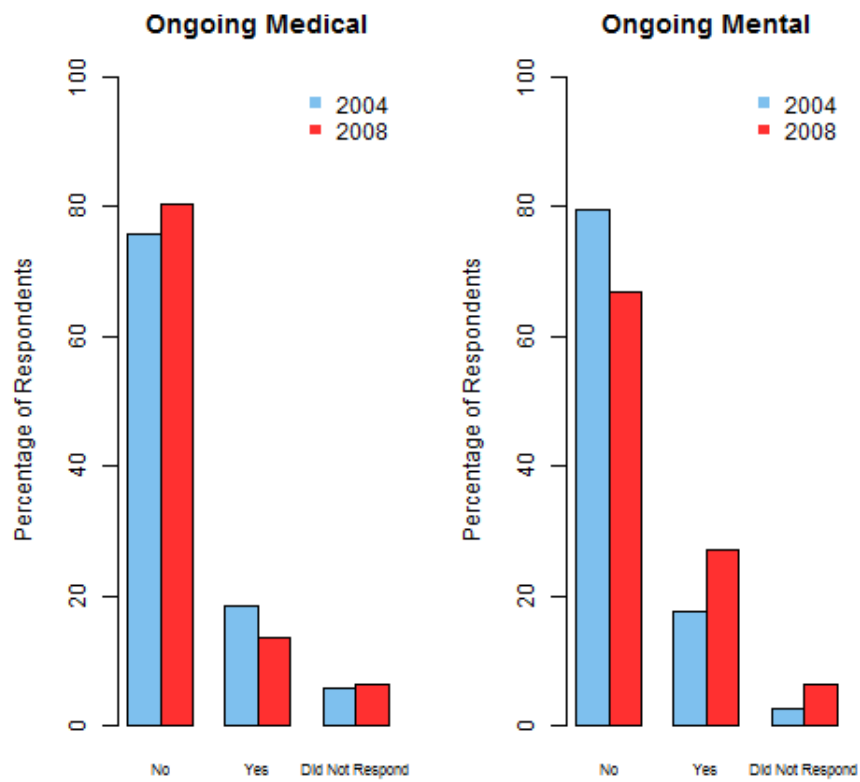


Fig. 2.12: Distribution of if individual has ongoing health condition since becoming clean (2004/2008)(unweighted)

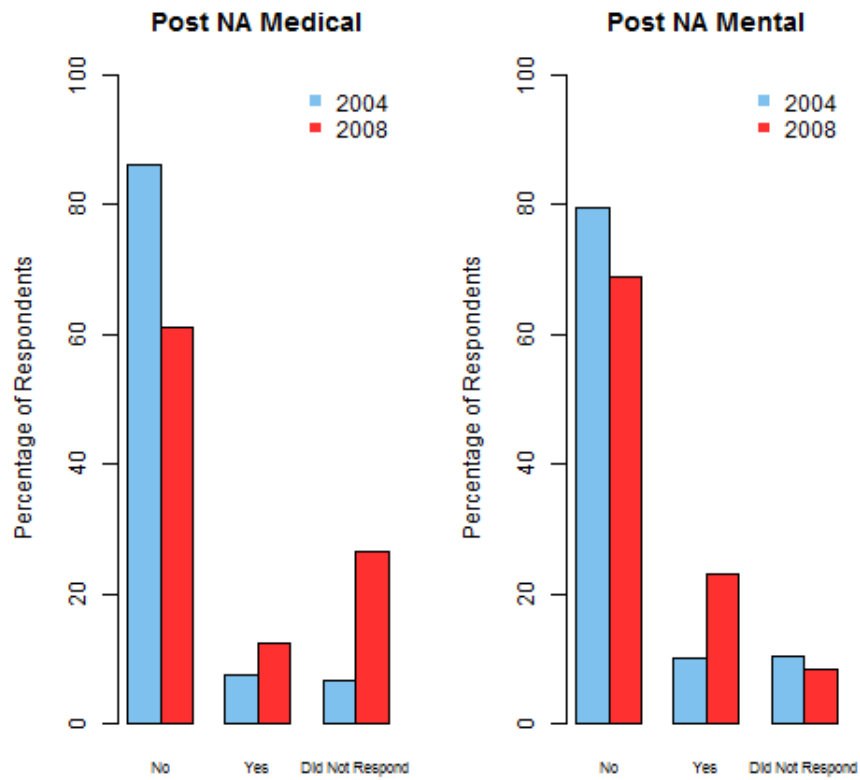


Fig. 2.13: Distribution of individual's health condition after joining NA (2004/2008)(unweighted)

Both surveys had 56% of respondents having criminal records due to their drug use, as shown in Figure 2.14.

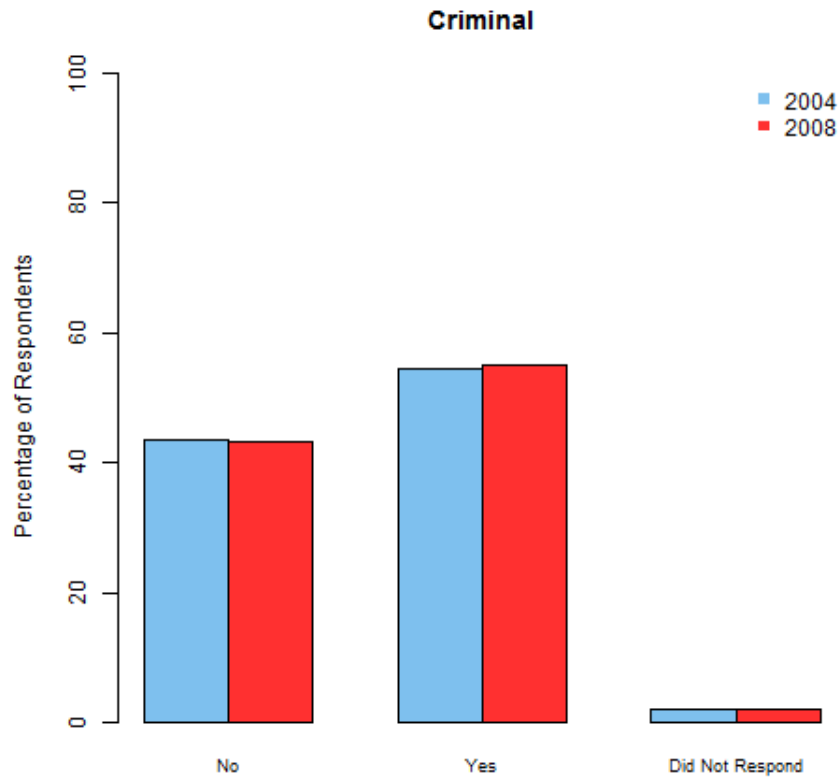


Fig. 2.14: Distribution of if individuals holds a criminal record due to drug use (2004/2008)(unweighted)

With the data source and survey designed defined, cross sectional estimation of the population size for Narcotics Anonymous can be established.

3. CROSS-SECTIONAL ESTIMATION

This section develops the previous theory established by Richard Arnold and Sharleen Forbes (Arnold & Forbes 2005). This includes the sample design, weight development, and population size estimates.

3.1 *Cross-Sectional Estimation Methodology*

3.1.1 *Sample Design*

There are two levels to the methodology. The first looks at the meeting level, while the second looks at the individual level.

There are M meetings of which m respond, we define the indicator variable as:

$$K_j = \begin{cases} 1, & \text{if meeting participated} \\ 0, & \text{otherwise} \end{cases} \quad \text{for } j = 1, \dots, M \quad (3.1)$$

Using this indicator variable, the number of responding meetings is:

$$m = \sum_{j=1}^M K_j, \quad (3.2)$$

and the meeting response rate is:

$$\phi_m = \frac{m}{M} = \frac{\# \text{ Meetings Participated}}{\# \text{ Number of Meetings Held}}. \quad (3.3)$$

An essential key to the design of this methodology is at the participating meetings a scribe was nominated, this scribe's job was to collect information on number attending the meeting, A_j , number completing the survey, C_j , and number who have previously completed a survey at an earlier meeting, H_j .

Using this information, the number of individuals who refused to participate in the survey is:

$$Z_j = A_j - C_j - H_j. \quad (3.4)$$

Due to the fact that all meetings were not forced to participate, values of A_j , H_j and Z_j are unknown for non-responding meetings though we do have $C_j = 0$ for those meetings. It becomes necessary now to impute the values A_j , H_j , and Z_j for the meetings that refused to answer. There are various ways of doing this, Richard Arnold's and Sharleen Forbes' paper "A method for estimating active membership for organisations that do not maintain registers of members", established two methods.

In the paper the two methods that were discussed;

- For small sized non-responding meetings an estimate of A_j (\hat{A}_j) can be made by speaking to individuals who attended those meetings during the survey week, and asking them to recall/estimate A_j
- For large non-responding meetings, or when no information is available from individual members, then a standard imputation approach can be used. In the paper a median meeting size for category c was defined by day and urban/rural status ($D \times V$). The size of non-responding meetings is imputed to be the median meeting size in the category.

Next we need a method for estimating the number H_j of previously completed surveys.

Due to nature of meetings refusing to participate individuals at these meetings have not completed the form i.e. $C_j = 0$. In order to estimate the number of previously completed individuals H_j , two methods were discussed,

- Assuming the people who attend refusal meetings will not respond in any meeting they attend $\hat{H}_j = 0$;
- Imputation can be done assuming that the proportion who have already completed at that meeting will be the same as another meeting held at the same time, same day, or based on another category.

Adopting the notation used in Arnold's and Forbes' paper (Arnold & Forbes 2005). There are N members of the population in which $i = 1, \dots, N$. In any given week there is:

$$A_{ij} = 1 \quad \text{if person } i \text{ attends meeting } j. \quad (3.5)$$

Extending this to the individual level, individual i attends $B_i = \sum_{j=1}^M A_{ij}$

meetings. At the meeting level there are $A_j = \sum_{i=1}^N A_{ij}$, individuals attending meeting j .

There are N_S distinct individuals that attend meetings during a given week, with each member of the population having a probability p_{ij} of attending a meeting in a given week. Using this probability p_{ij} the expected number of meetings attended by individual i in any given week is:

$$g_i = \sum_{j=1}^M p_{ij}. \quad (3.6)$$

The probability of attending meeting j is ≥ 0 ; in turn the expected number of meetings attended by individual i is $g_i \geq 0$. If $g_i = 0$ this means individual i does not ever attend any meetings and therefore is not part of our target population, therefore $g_i > 0$ is assumed.

Because the survey ran in just the survey week, the number of distinct individuals who attended, N_S , will be less than the entire population N . This is due to some individuals who attend meetings infrequently (fortnightly, monthly) and who may be absent during the survey week. These individuals are considered to be part of the population, N , but not included in the sample, N_S .

3.1.2 Weights

The question of the individual's frequency of attending meeting was asked in the survey. Because an individual is able to attend less than one meeting per week on average, some weighting is required. Individual i attends g_i meetings per week. The probability that an individual is captured in the survey week is worked out by taking the minimum of 1 or their meeting frequency g_i . We have this probability as if an individual attends fortnightly the probability of being captured in the survey week is $\frac{1}{2}$. Thus, we write the probability of being selected $\pi_i = \min(1, g_i)$.

We use this probability and define the (Horvitz-Thompson) inverse probability weight:

$$w_i = \frac{1}{\pi_i} = \frac{1}{\min(1, g_i)}. \quad (3.7)$$

The inverse probability weight defines how many individuals a single respondent represents of the population.

If an individual attends 4 times in a week they are expected to be captured 3 extra times during the survey week, but they still represent the same individual. We have the weight of $w_i = \frac{1}{\min(1,4)}$. However, if an individual attends once every 4 weeks, the assumption is made that there is another 3 individuals that attend meetings with the same frequency. These extra 3 individuals are not picked up in the survey week. The assumption is also made that these individuals share the same characteristics. Thus, the individual who attends every 4 weeks is given the weighting of $w_i = \frac{1}{\min(1,0.25)} = \frac{1}{0.25} = 4$.

3.1.3 Point Estimates

Using the information about the number of individuals present;

$$Z = \sum_{j=1}^M Z_j = \# \text{ attendances by non-compliant individuals;} \quad (3.8)$$

$$A = \sum_{j=1}^M A_j = \# \text{ attendances in the survey week;} \quad (3.9)$$

$$C = \sum_{j=1}^M C_j = \# \text{ of survey forms collected;} \quad (3.10)$$

$$H = \sum_{j=1}^M H_j = \# \text{ completed at previous meeting.} \quad (3.11)$$

$$\hat{Z} = \hat{A} - C - \hat{H}, \quad (3.12)$$

where $j = 1, 2, 3, \dots, M$, this can be written in such a way that, the individuals who have completed the form and the attendance of who refused to answer are known by:

$$\hat{Z} + C = \hat{A} - \hat{H}. \quad (3.13)$$

Next we need to know the proportion of individuals whose attendance is their first attendance. We assume that this proportion can be estimated from the compliant population:

$$\hat{\lambda} = \frac{C}{C + \hat{H}}, \quad (3.14)$$

where $C + \hat{H}$ is the total number of attendances by the individuals who completed the form at some point during the survey week. We now represent the number of non compliant individuals as:

$$\begin{aligned}
 \tilde{Z} &= \frac{C}{C + \hat{H}} \hat{Z} \\
 &= \frac{C}{C + \hat{H}} (\hat{A} - C - \hat{H}) \\
 &= \frac{C}{C + \hat{H}} (\hat{A} - (C - \hat{H}) - \hat{H}) \\
 &= \frac{C}{C + \hat{H}} (\hat{A} - C) \\
 &= \frac{C}{C + \hat{H}} \hat{A} - \frac{C}{C + \hat{H}} C \\
 &= \frac{C}{C + \hat{H}} \hat{A} - \frac{C}{C + \hat{H}} (C + \hat{H}) \\
 &= \frac{C}{C + \hat{H}} \hat{A} - C.
 \end{aligned}$$

The point estimate of the number of individuals that attend meetings each week is:

$$\hat{N}_S = C + \tilde{Z} = \frac{C}{C + \hat{H}} \hat{A} = \frac{C}{\hat{\phi}_c}, \quad (3.15)$$

where

$$\hat{\phi}_c = \frac{C + \hat{H}}{\hat{A}}, \quad (3.16)$$

which is the proportion of attendance that are compliant people, which is the estimate of proportion of the population that is compliant.

The point estimate is based on the assumption, that attendance patterns are the same for compliant and non-compliant individuals.

3.1.4 Confidence Intervals

The lower bound on the estimate N_S , is given by assuming that everyone in the population attends the meetings and no one refuses to answer the survey.

This means the number of forms collected is the size of N_S . As a result:

$$N_{S;l} = \sum_j C_j = C. \quad (3.17)$$

is the lower bound of the estimate N_S .

The upper bound on the estimate N_S , is the count of all attendance, treating every attendance as a distinct person. Because there are non responding meetings we use estimated values for these meetings and have the upper bound of N_S as:

$$N_{S;u} = \sum_{ij} A_{ij} = \sum_j \hat{A}_j = \hat{A}. \quad (3.18)$$

However, this estimate of the upper bound includes individuals who attend more than one meeting and have previously answered the survey, we adjust the upper bound by removing the number already known to have completed, thus, the upper bound becomes:

$$\begin{aligned} N_{S;u} &= \sum_j (A_j - H_j) \\ &= \hat{A} - \hat{H} \\ &= C + \hat{A} - C - \hat{H} \\ &= C + \hat{Z}. \end{aligned}$$

This treats all non-compliant attendances as distinct individuals. These estimates are used to gain insight on the estimate of the total population size. To accomplish this, the estimate N_S is divided by the proportion of the population that attends meetings in any given week. In order to do this the proportion of the population that attends meetings in any given week must be established. This is done using the formula:

$$\hat{\Psi} = \frac{\# \text{ Respondents}}{\# \text{ individuals respondent represents}} = \frac{\sum_{j=1}^M C_j}{\sum_{i \in s} w_i} = \frac{C}{W}, \quad (3.19)$$

where s is the responding population, and $W = \sum_{i \in s} w_i$ is the estimated size of the compliant population.

The point estimate of the total population is then:

$$\begin{aligned} \hat{N} &= \frac{\hat{N}_S}{\hat{\Psi}} \\ &= \frac{\frac{C}{C + \hat{H}} \hat{A}}{\frac{C}{W}} \\ &= \frac{W \hat{A}}{C + \hat{H}}, \end{aligned} \quad (3.20)$$

extending this argument into our error bounds we have:

$$N_l = \frac{C}{\widehat{\Psi}}, \quad (3.21)$$

and

$$\begin{aligned} N_u &= \frac{C + \widehat{Z}}{\widehat{\Psi}} \\ &= \frac{\widehat{A} - \widehat{H}}{\widehat{\Psi}} \\ &= \frac{W(\widehat{A} - \widehat{H})}{C}. \end{aligned} \quad (3.22)$$

3.1.5 Improved Estimates

An individual is only present if they are in the NA population at the time of the survey. If an individual is not present this implies they are not attending, not compliant, and not responding. An individual is responding if and only they are attending and are compliant.

$$I_i = \begin{cases} 1, & \text{if individual } i \text{ is present} \\ 0, & \text{Otherwise} \end{cases} \quad (3.23)$$

This population is broken down into two subgroups;(1) those who attend meetings during survey week, and (2) those who are absent from the meetings during the survey week.

$$a_i = \begin{cases} 1, & \text{if individual } i \text{ is attending meeting} \\ 0, & \text{Otherwise} \end{cases} \quad (3.24)$$

From those attending the meeting an individual may comply to complete a survey, or they can refuse to participate.

$$v_i = \begin{cases} 1, & \text{if individual } i \text{ is compliant} \\ 0, & \text{Otherwise} \end{cases} \quad (3.25)$$

If an individual is compliant this determines if the individual is responding. Responding individuals are defined as

$$r_i = \begin{cases} 1, & \text{if individual } i \text{ is responding} \\ 0, & \text{Otherwise} \end{cases} \quad (3.26)$$

A responding individual must be present in the population, attending, and compliant, such that:

$$r_i = I_i a_i v_i. \quad (3.27)$$

Recall from Section 3.1.2, the probability that a responding individual attends is:

$$\pi_i = \min(1, g_i), \quad (3.28)$$

with the (Horvitz-Thompson) inverse probability weights of:

$$w_i = \frac{1}{\pi_i} = \frac{1}{\min(1, g_i)}. \quad (3.29)$$

The sum of the weights of the respondents which estimates the size of the compliant population:

$$W = \sum_{i \in S} w_i. \quad (3.30)$$

The proportion of individuals whose attendance is their first attendance at time is:

$$\lambda = \frac{C}{C + \widehat{H}}. \quad (3.31)$$

The proportion of the population that attends meetings in any given week is:

$$\Psi = \frac{C}{W}. \quad (3.32)$$

The assumption of the attendance probability is the same for complying individuals and non-complying individuals. The proportion of the population that attends and is compliant is:

$$\phi_c = \frac{\text{\#individuals attending and are compliant}}{\text{\#individuals attending}} = \frac{C + H}{A}. \quad (3.33)$$

The estimated size of the population is:

$$\widehat{N} = \frac{W \widehat{A}}{C + \widehat{H}} = \frac{W}{\phi_c}. \quad (3.34)$$

3.1.6 Improved Confidence Interval

To estimate the standard error of \hat{N} , a synthetic population set U^* is generated and observed. The synthetic population is made of a compliant set and a non-compliant set. The compliant set is generated by generating $w_i = \frac{1}{\pi_i} = \frac{1}{\text{Prob. of attending}}$ copies of all individuals i in the respondent set s_i . All these generated individuals have $v_i^* = 1$.

When an individual has a non-integer w_i these individuals are generated by first generating a number of copies equal to the integer part of w_i . Then an additional copy is generated for individual i with a probability of $w_i - \text{int}(w_i)$.

The non-compliant set is generated by taking $w_i\delta$ copies of individual i , where δ is the odds of being non-compliant.

$$\delta = \frac{1 - \phi_c}{\phi_c} = \frac{1}{\phi_c} - 1 = \frac{A}{C + \hat{H}} - 1.$$

All these generated individuals have $v_i^* = 0$.

Using this synthetic population, the synthetic attendance for each individual a_i^* is also generated. The synthetic attendance is generated by using the probability of attendance, π_i^* , where $a_i^* \sim \text{bernoulli}(\pi_i^*)$.

The synthetic estimates are:

$$\begin{aligned} C^* &= \sum_{i \in U^*} r_i^*; \\ A^* &= \sum_{i \in U^*} a_i^* \max(1, g_i^*); \\ H^* &= \sum_{i \in U^*} a_i^* \max(1, g_i^*) v_i^* - C^*; \\ W^* &= \sum_{i \in U^*} w_i^*. \end{aligned}$$

Using equation 3.34 the synthetic population estimate is:

$$\hat{N}^* = \frac{W^* A^*}{C^* + H^*} \quad (3.35)$$

This synthetic population is repeatedly generated, and the standard deviation of the synthetic population estimate is an estimate of the standard error of \hat{N} .

3.2 Results

3.2.1 2004 Population Size Estimate

There were 85 meetings held during the survey week in 2004 (M). There were 3 group refusals leaving 82 groups participating (m). This gives the meeting response rate:

$$\phi_m = \frac{82}{85} = 96\%.$$

There are a total of $A = 1142$ individuals attending meetings, after imputation the total number of individuals increase to $\hat{A} = 1172$, $C = 475$ individuals completed the survey, with an additional $\hat{H} = 422$ individuals attending meetings but having already completed the survey at a previous meeting.

Recall from equation 3.15 the point estimate of the number of people attending meetings each week is:

$$\hat{N}_S = \frac{C}{C + \hat{H}} \hat{A}.$$

Thus we get our point estimate of the number of people attending meetings each week as:

$$\hat{N}_S = \frac{C}{C + \hat{H}} \hat{A} = \frac{475}{475 + 422} 1172 = 620.62 = 621,$$

where:

$$\hat{N}_S \in (C, \hat{A} - \hat{H}).$$

Thus the point estimate of N_S lies between:

$$(475, 1172 - 422) = (475, 750).$$

The proportion of attendances that are their first attendance is; the number of individuals completed divided by the number completed plus the number of individuals who have already completed at a previous survey. This becomes:

$$\lambda = \frac{475}{475 + 422} = \frac{475}{897} = 52.95\%.$$

As a result we have the response rate of individuals as:

$$\hat{\phi}_c = \frac{C}{\hat{N}_S} = \frac{475}{621} = 76.49\%.$$

The estimation of the proportion of the population that attends one or more meetings in any given week is:

$$\hat{\Psi} = \frac{C}{W} = \frac{475}{532.6} = 89.19\%.$$

The proportion of the population that attends meetings in any given week now becomes:

$$\hat{N} = \frac{\hat{N}_S}{\hat{\Psi}} = \frac{621}{89.19\%}.$$

We end up with the interval estimate of the total population as:

$$N \in \left(\frac{C}{\hat{\Psi}}, \frac{\hat{A} - \hat{H}}{\hat{\Psi}} \right) = \left(\frac{475}{89.19\%}, \frac{1172 - 422}{89.19\%} \right) = (533, 841).$$

The full breakdown of the frequency of attendance can be found in the Table 3.1.

Tab. 3.1: Frequency of attendance and weights of 2004 survey population

Meeting Frequency	Individuals	Sample Percentage	Weight w_i	Estimate $\sum_i w_i$	Weighted Percentage
At Least one Weekly	443	93.2%	1	443	83.2%
4 times per month	4	0.8%	1.1	4.3	0.8%
3 times per month	5	1.1%	1.4	7.2	1.4%
Twice per month	12	2.5%	2.2	26.0	4.9%
Once per month	10	2.1%	4.3	43.3	8.1%
Every two months	1	0.2%	8.7	8.7	1.6%
Total	475	100.00%		532.6	100.00%

We use the improved estimation method to improve these estimates. The improved standard error of the population estimate is 24.46, and the improved confidence interval for the 2004 estimate is $695 \pm 1.96 \times 24.46$, with the estimate of $\hat{N} = 695$ having a confidence interval of (647.05, 742.95).

3.2.2 2008 Population Size Estimate

Looking at the group level, in 2008 there were 90 meetings held during the survey week (M). There were 4 group refusals; leaving 86 groups participating (m). This gives the meeting response rate:

$$\phi_m = \frac{86}{90} = 95.56\%$$

The total of number of individuals attending is $A = 1208$, after imputation there is an estimated number of individuals attending $\hat{A} = 1256$ with $C = 546$ individuals completing the survey, with an additional $\hat{H} = 485$ individuals attending having already completed the survey at a previous meeting.

Recall equation 3.15 the point estimate of the number of people attending meetings each week is:

$$\hat{N}_S = \frac{C}{C + \hat{H}} \hat{A}.$$

The point estimate of the number of people attending meetings each week is:

$$\hat{N}_S = \frac{C}{C + \hat{H}} \hat{A} = \frac{546}{546 + 485} 1256 = 665.16 = 665,$$

where:

$$N_S \in (C, A - H).$$

The point estimate of N_S lies between:

$$(546, 1256 - 485) = (546, 771).$$

The proportion of attendances that are their first attendance is the number of individuals completed divided by the number completed plus the number of individuals who have already completed at a previous survey. We have:

$$\lambda = \frac{546}{546 + 485} = \frac{546}{1031} = 52.96\% = 53\%.$$

As a result we the response rate of individuals is:

$$\hat{\Phi}_c = \frac{C}{\hat{N}_S} = \frac{546}{665} = 82.1\%.$$

The estimation of the proportion of the population that attends one or more meetings in any given week is:

$$\Psi = \frac{C}{W} = \frac{546}{593.06} = 92\%.$$

The population that attends meetings in any given week now becomes:

$$\hat{N} = \frac{\hat{N}_S}{\hat{\Psi}} = \frac{665}{92\%} = 771.1 = 771.$$

Tab. 3.2: Frequency of attendance and weights of 2008 survey population

Meeting Frequency	Individuals	Sample Percentage	Weight w_i	Estimate $\sum_i w_i$	Weighted Percentage
At Least one Weekly	523	95.78%	1	523	88.19%
4 times per month	2	0.37%	1.08	2.17	0.37%
3 times per month	5	0.92%	1.44	7.22	1.21%
Twice per month	4	0.73%	2.17	8.67	1.46%
Once per month	12	2.20%	4.33	52	8.77%
Total	546	100.00%		593.06	100.00%

The resulting interval estimate of the total population is:

$$N \in \left(\frac{C}{\widehat{\Psi}}, \frac{\widehat{A} - \widehat{H}}{\widehat{\Psi}} \right) = \left(\frac{546}{92\%}, \frac{1256 - 485}{92\%} \right) = (592, 836).$$

The full breakdown of the frequency of attendance can be found in the Table 3.2

We use the improved estimation method to improve these estimates. The improved standard error of the size of the population estimate is 20.31, the improved confidence interval for the 2008 estimate is $771 \pm 1.96 \times 20.31$. With the estimate of $\widehat{N} = 771$ having a confidence interval of (731.20, 810.80).

3.3 Bootstrap Estimation

3.3.1 Bootstrap

There are multiple re-sampling techniques which include but are not limited by the Jackknife method and the Bootstrap method. This research uses the latter.

There are two main types of bootstrapping:

1. Non-parametric bootstrap - re-sample the observations from the sample gathered
2. Parametric bootstrap - build a theoretical model using estimated parameters, and re-sample from the distribution

The non-parametric bootstrap is the method of use and of interest in this research. From this point any reference to the bootstrap is to be assumed as the non-parametric bootstrap.

3.3.2 Bootstrap Methodology

- Draw B independent bootstrap samples based of the empirical distribution:

$$X_1^{*(b)}, X_2^{*(b)}, \dots, X_m^{*(b)},$$

- Evaluate:

$$\hat{\theta}_b^* = s(X_b^*),$$

- Estimate the point estimates of the B replications,

where $s(\cdot)$ is the function of the sample values which estimates θ . This is illustrated in Table 3.3

Tab. 3.3: Bootstrap methodology

	Data	Statistic
Sample	$\mathbf{x} = (x_1, x_2, \dots, x_j, \dots, x_m)$	$s(\mathbf{x})$
Bootstrap sample 1	$\mathbf{x}^{*1} = (x_1^{*1}, x_2^{*1}, \dots, x_j^{*1}, \dots, x_m^{*1})$	$s(\mathbf{x}^{*1})$
Bootstrap sample 2	$\mathbf{x}^{*2} = (x_1^{*2}, x_2^{*2}, \dots, x_j^{*2}, \dots, x_m^{*2})$	$s(\mathbf{x}^{*2})$
\vdots	\vdots	\vdots
Bootstrap sample b	$\mathbf{x}^{*b} = (x_1^{*b}, x_2^{*b}, \dots, x_j^{*b}, \dots, x_m^{*b})$	$s(\mathbf{x}^{*b})$
\vdots	\vdots	\vdots
Bootstrap sample B	$\mathbf{x}^{*B} = (x_1^{*B}, x_2^{*B}, \dots, x_j^{*B}, \dots, x_m^{*B})$	$s(\mathbf{x}^{*B})$

the sample values are denoted by $j = 1, 2, \dots, m$ and the bootstrapped samples by $b = 1, 2, \dots, B$

3.3.3 Point Estimates

The point estimate of the bootstrap estimate is:

$$\hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^* = \frac{1}{B} \sum_{b=1}^B s(\mathbf{x}^{*b}), \quad (3.36)$$

the standard error $se(\theta)$ is estimated by the standard deviation of the replications

$$\hat{\sigma}_{boot} = \sqrt{\frac{\sum_{b=1}^B (\theta_b^* - \hat{\theta}_{boot})^2}{B - 1}}. \quad (3.37)$$

3.3.4 Confidence Intervals

Once the point estimates of the sampling distribution have been established we can construct confidence intervals for these point estimates by using either standard confidence interval or percentile confidence intervals.

Standard confidence interval

Assume that the samples are normally distributed with mean and variance $N(\hat{\theta}_{boot}, \hat{\sigma}_{boot}^2)$. The lower bound is given by

$$\hat{\theta}_L = \hat{\theta}_{boot} - z_{\alpha/2} \hat{se}_{boot}(\hat{\theta}_{boot})$$

and upper bound is given by

$$\hat{\theta}_U = \hat{\theta}_{boot} + z_{\alpha/2} \hat{se}_{boot}(\hat{\theta}_{boot}),$$

where z_α is the α critical value. The 95% confidence interval is written as

$$(\hat{\theta}_L, \hat{\theta}_U) = (\hat{\theta}_{boot} - 1.96 \hat{se}_{boot}(\hat{\theta}_{boot}), \hat{\theta}_{boot} + 1.96 \hat{se}_{boot}(\hat{\theta}_{boot}))$$

This is the main interval which the analysis uses an alternative to this confidence interval is percentile interval.

Percentile Interval: The basic version of the percentile confidence interval is ordering the bootstrap estimates $\hat{\theta}_b^*$ and the $(1 - \alpha)$ confidence interval is defined by

$$\hat{\theta}_L = \alpha/2$$

and

$$\hat{\theta}_U = (1 - \alpha/2).$$

3.3.5 Bootstrap Estimation

The bootstrap estimates were done in two ways. The first is bootstrapping the unweighted sample estimates, the second bootstraps the weighted estimates. For each survey dataset, the data was re-sampled in a way that the same sized dataset was obtained, this was done based on the Horvitz-Thompson probability of $\frac{1}{w_i}$. An estimate was created this process was repeated 1000 times.

Once the 1000 estimates were established the mean of these estimates were taken to create the bootstrap estimate. The simple normal bootstrap confidence interval was used ($\hat{\theta} \pm 1.96se(\hat{\theta})$).

The variables that were bootstrapped were:

- Age;
- Clean Time;
- NA Time (association time);
- Education;
- Income Source;

Age, *Clean Time*, and *NA Time* are all numeric variables, where the mean is the estimation function used. While *Education* and *Income Source* are both categorical variables which require the proportion estimate as the estimation function.

Bootstrapping the 2004 data, the unweighed average age is 37 years old with an average clean time of 4.8 years and an average time in NA of 7 years. The weighted versions of these estimates don't change drastically, more over they have just one more year added to the estimates these variables on average.

Bootstrapping the 2008 data, the unweighted bootstrapped estimates give an average age of 39 years old, with an average clean time of 5 and a half years with the time in NA being 7.8 years on average. Employed is the highest income source, and a near even split for tertiary and graduate educations. The weighted bootstrapped estimates follow the same trends with a slight increase in the average clean time, age, and time in NA.

There is an increase in all the weighted estimates compared to the unweighted estimates. This means the individuals who are absent during the survey were, older individuals who had been in the programme for longer and had a longer clean time. This is not unexpected as individuals who have remained in NA

Tab. 3.4: Table of 2004 Bootstrap Estimates

Variable	level	Unweighted Estimate	CI	Weighted Estimate	CI
Age		37.27	(37.26, 37.28)	38.74	(37.31, 40.18)
Clean Time		4.77	(4.76, 4.77)	5.03	(4.38, 5.68)
NA Time		7.04	(7.03, 7.04)	7.32	(6.64, 8.01)
Education	None	18.83%	(15.29, 22.37)	18.85%	(15.18, 22.52)
	School	21.21%	(17.46, 24.96)	21.31 %	(17.44, 25.18)
	Vocational	14.70%	(11.59, 17.81)	14.80%	(11.49, 18.10)
	Tertiary	26.41 %	(22.54, 30.29)	26.25%	(22.10, 30.40)
	Graduate	18.85%	(15.44, 22.26)	18.79%	(15.40, 22.19)
Income Source	Beneficiary	36.16%	(31.76, 40.59)	36.24%	(31.71, 40.78)
	Employed	53.74%	(49.24, 58.25)	53.62%	(49.02, 58.22)
	Not In Labour Force	5.39 %	(3.38, 7.41)	5.43%	(3.31, 7.55)
	Student	4.69%	(2.75, 6.62)	4.71%	(2.73, 6.68)

Tab. 3.5: Table of 2008 Bootstrap Estimates

Variable	level	Unweighted Estimate	CI	Weighted Estimate	CI
Age		39.53	(39.52, 39.54)	40.51	(39.28, 41.72)
CleanTime		5.43	(5.43, 5.44)	5.58	(4.96, 6.20)
NA Time		7.81	(7.81, 7.82)	8.08	(7.29, 8.86)
Education	None	21.31%	(17.83, 24.80)	21.32%	(17.72, 24.91)
	School	18.46%	(15.24, 21.68)	18.40%	(15.14, 21.66)
	Vocational	16.63%	(13.50, 19.76)	16.59%	(13.37, 19.80)
	Tertiary	23.26 %	(19.55, 26.98)	23.26%	(19.66, 26.85)
	Graduate	20.33%	(16.91, 23.75)	20.44%	(16.68, 24.19)
Income Source	Beneficiary	33.10%	(28.99, 37.22)	33.09%	(29.17, 37.01)
	Employed	60.93%	(56.77, 65.10)	60.94%	(56.77, 65.11)
	Not In Labour Force	4.08 %	(2.41, 5.75)	4.08%	(2.37, 5.80)
	Student	1.88%	(0.70, 3.06)	1.88%	(0.67, 3.10)

for longer would've established some coping mechanism which helps them deal with their addition compared to new members, and would require to attend less meetings.

Bootstrap estimation is used later in Section 3.5.1 on the regression models to estimate bootstrapped estimates of the regression model.

3.4 Statistical Theory

In addition to the simple summary statistics derived, other methods of statistical analysis are essential to analysis of this survey data. In addition to the bootstrap estimation regression analysis is also used.

Agresti (2007) summarised the type of analysis appropriate with the type of data in Table 3.6.

Tab. 3.6: Summary made by Agresti

Random Component	Explanatory	
	Variable	Model
Normal	Continuous	Regression
Normal	Categorical	Analysis of variance
Normal	Mixed	Analysis of covariance
Binomial	Mixed	Logistic regression
Poisson	Mixed	Loglinear
Multinomial	Mixed	Multinomial response

3.4.1 Weighted Point Estimates

For surveys where the probability of selection is known for each respondent the Horvitz-Thompson estimator is used.

The Horvitz-Thompson estimator for the mean of an estimate is:

$$\hat{Y}_{HT} = \frac{\sum_{k=1}^n w_k y_k}{\sum_{k=1}^n w_k} \quad (3.38)$$

where y_k is the value of the variable of interest e.g. Age, time, etc.

3.4.2 Regression

In regression analysis the type of model used depends on the dependent variable. When the dependent variable is continuous, linear regression can be used.

In this chapter only linear regression will be addressed, logistic regression is conducted in Chapter 6 in more detail.

General Linear Regression

When there are n pairs of continuous dependent variable Y and the continuous explanatory variables \mathbf{X} . The linear regression model takes the form:

$$y_i = \beta_0 + \sum_{j=1}^P \beta_j x_{ij} + \epsilon_{ij} \quad (3.39)$$

which can be modelled by:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad (3.40)$$

with the estimates of:

$$\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

3.5 Other Cross-Sectional Estimates

3.5.1 Linear Regression

Linear Regression on NA Time 2004

Starting with a look at the linear regression for the *NA Time* regressed on *Age*, *Sex*, *Ethnicity*, *Education*, and *Income Source* (Table 3.7). *Age* is grouped into the categories of:

- < 30 years old;
- 30 years old to but not including 45 years old;
- 45 years old to but not including 60 years old;
- 60 years or older.

Ethnicity has also been coded as a binary variable of either having Maori or Non-Maori. *Income Source* has been coded into “Employed”, “Beneficiary”, “Not in Labour Force” and “Student”. Both beneficiary and student were coded as their response, only the responses “Full Time Employment”, “Part Time Employment”, and “self employment” are encoded to the “Employed” factor. The rest are encoded to “Not In The Labour Force”.

Observing the data individuals that are a beneficiary have been in NA for 4 years less than individuals that are employed. For every 15 year increase the expected time in NA increases by approx 4 years up to the age of 60, there is no significant difference for sex and ethnicity.

Tab. 3.7: 2004 Bootstrapped regression NA Time

Coefficients	level	Estimate	Std. Error	p-value	
(Intercept)		2.43	1.00	0.015	*
Age	<30	0.00			
	30-(45)	4.11	0.68	<0.001	***
	45-(60)	7.89	0.83	<0.001	***
	60+	7.57	2.75	0.006	**
Sex	Male	0.00			
	Female	0.30	0.55	0.59	
Ethnicity	Non-Maori	0.00			
	Maori	0.02	0.74	0.979	
Education	None	0.00			
	School	1.63	0.91	0.071	.
	Vocational	0.87	1.01	0.391	
	Tertiary	2.79	0.86	0.001	**
Income	Graduate	2.25	0.92	0.015	*
	Employed	0.00			
	Beneficiary	-3.95	0.60	<0.001	***
	Not In				
	Labour	-0.96	1.32	0.468	
	Student	-1.82	1.19	0.124	

Linear Regression on Clean Time 2004

Regressing *Clean Time* against *Age*, *Sex*, *Ethnicity*, *Education*, and *Income Source*. This has a reference level of <30 year old Non-Maori Male, employed with no education, results are displayed in Table 3.8.

Tab. 3.8: 2004 Bootstrapped regression Clean Time

Coefficients		Estimate	Std. Error	p-value	
(Intercept)		2.37	0.81	0.004	**
Age	<30	0.00			
	30-(45)	2.12	0.58	<0.001	***
	45-(60)	5.47	0.71	<0.001	***
	60+	7.28	2.07	<0.001	***
Sex	Male	0.00			
	Female	0.87	0.47	0.065	.
Ethnicity	Non-Maori	0.00			
	Maori	0.28	0.63	0.651	
Education	None	0.00			
	School	0.93	0.76	0.223	*
	Vocational	0.68	0.82	0.408	
	Tertiary	1.82	0.72	0.012	*
Income	Graduate	1.78	0.78	0.022	
	Employed	0.00			
	Beneficiary	-4.21	0.51	<0.001	***
	Not In				
	Labour	-2.02	1.10	0.019	*
	Student	-2.51	1.07	0.068	.

Individuals that are a beneficiary have a clean time of 4 years lower than individuals that are employed, for every 15 year increase the expected clean time increases doubles until the age of 60, there is no significant difference for sex and ethnicity.

Linear Regression on NA Time 2008

Regressing *NA Time* in 2008 against *Age*, *Sex*, *Ethnicity*, *Education*, and *Income Source*. This again has a reference level of <30 year old Non-Maori Male, employed with no education. The outcome of the regression model is shown in Table 3.9.

Tab. 3.9: 2008 Bootstrapped regression NA Time interaction

Coefficients	levels	Estimate	Std. Error	p-value	
(Intercept)		2.78	1.08	0.010	*
Age	<30	0.00			
	30-(45)	3.42	0.84	<0.001	***
	45-(60)	8.31	0.93	<0.001	***
	60+	9.59	1.95	<0.001	***
Sex	Male	0.00			
	Female	0.47	0.65	0.473	
Ethnicity	Non-Maori	0.00			
	Maori	1.53	0.88	0.083	.
Education	None	0.00			
	School	1.15	1.06	0.276	
	Vocational	1.45	1.07	0.175	
	Tertiary	1.15	0.97	0.238	
	Graduate	2.18	1.00	0.031	*
Income	Employed	0.00			
	Beneficiary	-3.81	0.70	<0.001	***
	Not In				
	Labour	-1.43	1.67	0.392	
	Student	-3.08	2.30	0.1811	

Individuals that are a beneficiary have a been in NA 4 years less than individuals that are employed. For every 15 year increase the expected time in NA increases by roughly doubles until age 60. There still is no significant difference for sex and ethnicity.

Linear Regression on Clean Time 2008

Tab. 3.10: 2008 Bootstrapped regression Clean Time

Coefficients		Estimate	Std. Error	p-value	
(Intercept)		1.84	0.88	0.037	*
Age	<30	0.00			***
	30-(45)	2.36	0.69	0.001	***
	45-(60)	5.87	0.77	<0.001	***
	60+	7.16	1.63	<0.001	***
Sex	Male	0.00			
	Female	-0.49	0.55	0.376	
Ethnicity	Non-Maori	0.00			
	Maori	0.69	0.71	0.335	
Education	None	0.00			
	School	0.90	0.87	0.301	
	Vocational	1.42	0.87	0.105	
	Tertiary	2.04	0.83	0.014	*
Income	Graduate	3.62	0.84	<0.001	***
	Employed	0.00			
	Beneficiary	-3.60	0.59	<0.001	***
	Not In				
	Labour	0.04	1.38	0.979	
	Student	-2.87	1.82	0.114	

There are many similarities between the regression models of 2004 clean time (Table 3.8) and 2008 clean time (Table 3.10), which the same conclusions about individuals doubling their clean time for every 15 years of ageing, and the beneficiaries have a significantly lower clean time than the employed individuals.

The model for interaction was tested for each model. In each case the interaction between *Age*, *Sex*, *Ethnicity*, and *Income Source* had no effect.

3.6 Concluding Remarks

From the linear regression both time in NA and clean time are expected to be longer as the individual ages. This is showing the older an individual is the longer the expected clean time/time in NA is compared to a younger individual.

There is also significant differences in the clean time and time in NA for individuals who are employed and those who are beneficiaries. With individuals who are employed showing longer clean time/time in NA than their beneficiary counterpart.

In all four models individuals who obtained a tertiary qualification (Graduate or postgraduate qualification) showed significant differences to individuals without any qualification.

This establishes the cross-sectional information. The longitudinal estimates are developed next. The first part of longitudinal analysis is probabilistic matching, the theory is developed in chapter 4.

4. RECORD LINKAGE THEORY

In the attempt of linking two records together, there are two distinct ways of linking (matching) these records together. The first is a deterministic record linkage approach, and the second is a probabilistic record linkage approach. The need for a record linkage arises from either:

- The construction/ maintenance of a master file for a population;
- Merging two files to aid in extending the amount of information about the population represented in both files.

In any record linkage study there are at least two population datasets A and B , where individual elements are denoted by a_i and b_j .

There is a key assumption that is made in any record linkage. This assumption is that there is at least 1 element of A that is also an element of B . This assumption is made as without at least 1 element in common there is no reason to try to link the two datasets.

The set that results from linking every record in A to every record in B is:

$$A \times B = \{(a_i, b_j) : a_i \in A, b_j \in B\} \quad (4.1)$$

which can be broken down into three separate cases:

1. Records in A that do not match any records in B ;
2. Records in A that match at least one record in B ;
3. Records in B that do not match any record in A .

Each comparison (a_i, b_j) may or may not be a match.

As a result the set $A \times B = \{(a, b) : a \in A, b \in B\}$ is written as the union of the two disjoint sets:

$$M = \{(a_i, b_j) : a_i = b_j, a_i \in A, b_j \in B\} \quad (4.2)$$

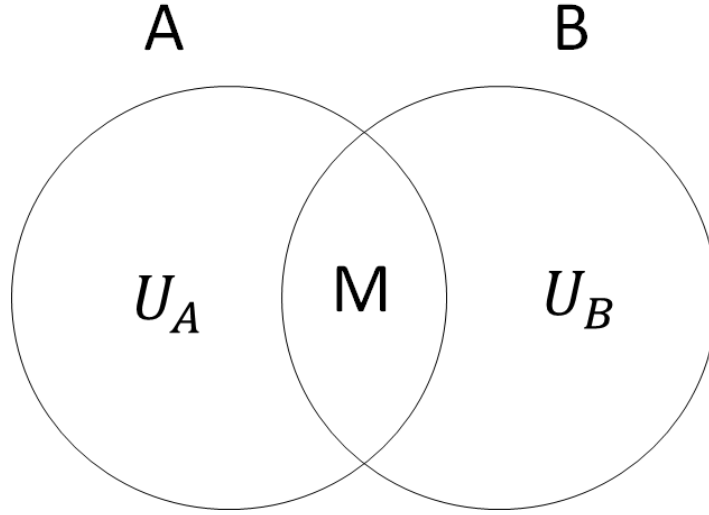


Fig. 4.1: Illustration of matched and unmatched sets

and

$$U = \{(a_i, b_j) : a_i \neq b_j, a_i \in A, b_j \in B\} \quad (4.3)$$

which are called the matched and unmatched sets respectively. Two individuals (a_i, b_j) are considered a matched pair if they are in the set M and are unmatched pairs (do not match) if they are in the set U (See Figure 4.1). Duplicate matches can occur, this is when an individual a_i is matched to more than one individual b_i in B .

4.1 Deterministic Matching

Deterministic Record Linkage, referred to as deterministic matching, is one of the most simple forms of Record Linkage. The deterministic method matches the different datasets by a unique identifier which is held in common to both datasets. A record is said to be a match if the deterministic matching successfully identifies the same individual using some if not all, unique identifiers.

Deterministic matching is only possible when each data set shares a least one *unique* identifier between the two data sets. It is possible to extend the idea behind deterministic record linkage from two datasets, to three or more datasets.

Examples of deterministic matching are common place ranging from finance to medical research.

Example

Suppose we wish to conduct a study to find the employment rate of graduates from a particular university. The university is unable to see what happens to the graduate when they finish their studies, but once the student has left the university the IRD have knowledge of them in the workforce, at registration the university realises this and asks for the student's IRD number.

Suppose we have the following dataset layouts:

Tab. 4.1: University Record

Student Id	IRD Number	Status
123456789	987-654-32	Graduate

Tab. 4.2: IRD Record

Id	IRD Number	Status
1	987-654-32	Full-Time

where the university doesn't know the students work status, nor does the IRD know the individuals education status. Both datasets share the IRD number in common, as a result a deterministic matching can be conducted using this as a unique identifier, and we can find out the employment status of the student using this deterministic matching.

Suppose we want to know the employment status of only graduates, first we refine the university record to have the status of "Graduate" only, using this subgroup we then match the IRD number of these students to the IRD from the IRD record, i.e. the student with a student id 123456789 is a graduate with the IRD number 987-654-32, which is deterministically match to record 1 which states that this individual is in Full-Time employment.

In the recording process of the datasets, some errors are potentially introduced. These errors can be caused by errors in the data entry process e.g. errors in coding, transcribing, keypunching etc. Due to these errors some of the true matches (a correct matching between an individual in dataset A and an individual in dataset B) will be missed. On rare occasions these errors can introduce false matches (an incorrect matching of an individual in dataset A and an individual in dataset B).

Due to situations where the data has been incorrectly entered or been left empty, or when data sets do not have a unique identifier in common deterministic record linkage is insufficient as a matching method. This insufficiency led to the rise of the theory for probabilistic matching.

4.2 Probabilistic Matching

Most of the early work in probabilistic record linkage was conducted by Fellegi and Sunter, in their 1969 paper A Theory for Record Linkage (Fellegi & Sunter 1969). The Fellegi-Sunter Model uses a decision-theoretic approach which establishes the validity of principles, first used in practice by Newcombe (Newcombe et al. 1959).

In the process of Probabilistic Record Linkage, referred to as Probabilistic Matching, there are two main cases that must be also considered:

- The case where all variables are treated as independent.
- The case where all variables are treated as dependent.

In this research only the first case is investigated.

For the dataset A the variable a_{ik} is defined as the i^{th} entry on data set A for the k^{th} variable. Similarly for dataset B the variable b_{jk} is defined as the j^{th} entry on data set B for the k^{th} variable. As a result we define the vector function of:

$$\gamma_{ij} = (\gamma_1(a_{i1}, b_{j1}), \dots, \gamma_k(a_{iK}, b_{jK})). \quad (4.4)$$

This is the comparison vector of comparing data set A and data set B where;

$$\gamma_k(a_{ik}, b_{jk}) = \begin{cases} 1 & \text{if } a_{ik} \text{ agrees with } b_{jk} \\ 0 & \text{otherwise} \end{cases} \quad (4.5)$$

The comparison vector is also written as $\gamma_{ij} = \gamma(a_i, b_j)$. The entire comparison space is denoted by Γ .

4.2.1 Slack

Two individuals a_i and b_j will agree to be a match if they are within some specified “slack” amount of each other. The two individuals a_i and b_j agree if numerical variable k has values satisfying:

$$|a_{ik} - b_{jk}| \leq \delta_k. \quad (4.6)$$

This δ_k is the slack for the k^{th} variable. δ_k is a pre-determined numeric value for all the numeric variables and 0 for all the categorical variables.

Individuals a_i belong to the dataset A with $|A| = n_A$, and individuals b_i belong to the dataset B with $|B| = n_B$. The set of all comparison vectors is the comparison space which covers the entire space $|A \times B|$ and is defined by:

$$\gamma_{ijk} = \gamma(a_i k, b_j k) \quad i = 1, \dots, n_A, \quad j = 1, \dots, n_B \quad k = 1, \dots, K.$$

Individuals are considered a match based on weights, these weights are calculated by using this comparison space, u -probabilities and m -probabilities which are derived in the sections 4.2.2 and 4.2.3.

4.2.2 u -probability

The u -probabilities are defined as:

$$u_k = \Pr(a_i \text{ and } b_j \text{ agree on variable } k | a_i \text{ and } b_j \text{ are not a true match}).$$

We write u_k as:

$$u_k = \Pr(\gamma_{ijk} = 1 | U_{ij} = 1), \quad (4.7)$$

where

$$u_k = \Pr(\text{a randomly chosen pair of individuals } (a_i, b_i) \in U \\ \text{match on variable } k)$$

For example if variable k is a binary variable then:

$$u_k = \Pr(a_{ik} = 1 \text{ and } b_{jk} = 1 | \text{not a true match}) \\ + \Pr(a_{ik} = 0 \text{ and } b_{jk} = 0 | \text{not a true match}).$$

As we are assuming the variables are independent and matches are rare as a result we write:

$$u_k = \Pr(a_{ik} = 1)\Pr(b_{jk} = 1) + \Pr(a_{ik} = 0)\Pr(b_{jk} = 0) \\ = p_A p_B + (1 - p_A)(1 - p_B),$$

where $p_{Ak} = \text{Prob}(A_{ik} = 1)$ and $p_{Bk} = \text{Prob}(B_{jk} = 1)$.

A simple estimate of the u -probability is the proportion matched in the comparison space, Γ , to the size of Γ .

$$u_k = \frac{\sum_i \sum_j \gamma_{ijk}}{n_A n_B} \quad (4.8)$$

Example

Consider two simulated datasets, A and dataset B, which share 5 variables:

- First Name,
- Last Name,
- Day of Birth,
- Month of Birth,
- Year of Birth.

Dataset A contains 200 individuals ($|A| = n_A = 200$), and dataset B contains 250 individuals ($|B| = n_B = 250$).

These two datasets have 30 entries in common. In order to determine which entries are in common a probabilistic match is conducted.

In order to complete a probabilistic matching first we must determine the u -probabilities. From the comparison space we have $n_A n_B = 50,000$ entries for the each variable v_k . The slack that has been specified is:

$$\delta = (0, 0, 1, 0, 0),$$

i.e. we require exact match on all variables except day of birth, which has a slack of ± 1 day.

Using the formula:

$$u_k = \frac{\sum_i \sum_j \gamma_{ijk}}{n_A n_B}.$$

We get the u -probabilities of

$$\begin{aligned} u &= (u_1, u_2, u_3, u_4, u_5) \\ &= \left(\frac{\sum_i \sum_j \gamma_{ij1}}{200 * 250}, \frac{\sum_i \sum_j \gamma_{ij2}}{200 * 250}, \frac{\sum_i \sum_j \gamma_{ij3}}{200 * 250}, \frac{\sum_i \sum_j \gamma_{ij4}}{200 * 250}, \frac{\sum_i \sum_j \gamma_{ij5}}{200 * 250} \right) \\ &= \left(\frac{48}{50000}, \frac{91}{50000}, \frac{4683}{50000}, \frac{4188}{50000}, \frac{752}{50000} \right) \\ &= (0.00096, 0.00182, 0.09366, 0.08376, 0.01504). \end{aligned}$$

Knowing the u -probability is only part of the calculation for the weight used in probabilistic matching that needs to be done. In order to complete a probabilistic matching the m -probabilities are also required.

4.2.3 m -probability

The m -probability is defined as:

$$m_k = \Pr(a_i \text{ and } b_j \text{ agree on variable } k | a_i \text{ and } b_j \text{ are a true match}),$$

this simplifies to $m_k = \text{reliability of the variable } k \text{ as a matching variable}$.

When calculating the m -probability, a baseline estimate of the m -probability is created for each variable. From these estimates the number of matched individuals are calculated and as a result the empirical m_k probabilities are calculated. The estimated m_k probabilities are then set equal to the empirical m_k probabilities. This method is repeated until the m_k probabilities stabilises.

Now the m -probability has been derived in order to complete a probabilistic matching each variable needs a weight.

4.2.4 Weight

In the Fellegi-Sunter framework; the best record linkage rule involves calculating the agreement/disagreement weight. This overall weight is used to determine if two records are matched between the two sets, the higher the weight more likely the records match. The weight is composed of the sum of the individual variable weights.

We define $M_{ij} = 1$ if the pair ij is a genuine match, and $U_{ij} = 1 - M_{ij} = 1$ if the pair is not a true match.

Fellegi stated that any monotone increasing function of $\frac{m_k}{u_k}$ can be used to calculate the weights. However, in this paper we should stick to the basic form of taking logarithm of this ratio. Thus the weight calculated is:

$$w_k = \log \frac{m_k}{u_k} \quad (4.9)$$

Earlier we saw $m_k = \Pr(\gamma_{ijk} = 1 | M_{ij} = 1)$ and $u_k = \Pr(\gamma_{ijk} = 1 | U_{ij} = 1)$, thus we can write:

$$r_k = \frac{\Pr(M_{ij} = 1 | \gamma_{ij} = 1)}{\Pr(U_{ij} = 1 | \gamma_{ij} = 1)},$$

which is the odds of being a match if variable k matches. Recall Bayes Theorem:

$$Pr(A|B) = \frac{Pr(B|A)P(A)}{Pr(B)} \quad (4.10)$$

Thus we can write $Pr(M_{ij} = 1|\gamma_{ijk} = 1)$ as $\frac{Pr(\gamma_{ijk}=1|M_{ij}=1)Pr(M_{ij}=1)}{Pr(\gamma_{ijk}=1)}$.

We can also write $Pr(U_{ij} = 1|\gamma_{ijk} = 1)$ as $\frac{Pr(\gamma_{ijk}=1|U_{ij}=1)Pr(U_{ij}=1)}{Pr(\gamma_{ijk}=1)}$.

Using this information we can write the weights in the form of

$$w_k = \frac{Pr(\gamma_{ij} = 1|M_{ij} = 1)}{Pr(\gamma_{ij} = 1|U_{ij} = 1)} \times \frac{Pr(M_{ij} = 1)}{Pr(U_{ij} = 1)} \quad (4.11)$$

Since the value of γ_{ijk} can only take on values 0 and 1. Thus

$$\gamma_{ijk}|M_{ij} \sim \text{Bernoulli}(m_k)$$

and,

$$\gamma_{ijk}|U_{ij} \sim \text{Bernoulli}(u_k)$$

the weights can be written in the form of:

$$\prod_{k=1}^P \frac{m_k^{\gamma_{ijk}}(1-m_k)^{1-\gamma_{ijk}}}{u_k^{\gamma_{ijk}}(1-u_k)^{1-\gamma_{ijk}}} \frac{Pr(M_{ij} = 1)}{Pr(U_{ij} = 1)} \quad (4.12)$$

taking the logarithm of both sides the weights become

$$\begin{aligned} w_k &= \log \prod_{k=1}^P \frac{m_k^{\gamma_{ijk}}(1-m_k)^{1-\gamma_{ijk}}}{u_k^{\gamma_{ijk}}(1-u_k)^{1-\gamma_{ijk}}} + \log \frac{Pr(M_{ij} = 1)}{Pr(U_{ij} = 1)} \\ &= \sum_{k=1}^k \gamma_{ijk} \log \left(\frac{m_k}{u_k} \right) + \sum_{k=1}^P (1 - \gamma_{ijk}) \log \left(\frac{1 - m_k}{1 - u_k} \right) + \log \left(\frac{Pr(m_{ij})}{Pr(u_{ij})} \right) \end{aligned}$$

We can drop the term $\log \frac{Pr(M_{ij}=1)}{Pr(U_{ij}=1)}$ as this is a constant term across all weights, as a result the weights become:

$$w = \sum_{k=1}^P [\gamma_{ijk} A_k + (1 - \gamma_{ijk}) D_k], \quad (4.13)$$

where A_k is the agreement weight $\log \left(\frac{m_k}{u_k} \right)$ and D_k is the disagreement weight $\log \left(\frac{(1-m_k)}{(1-u_k)} \right)$.

Example Continued

Using the same datasets from earlier, the m -probabilities are set to, First Name = 0.90, Last Name = 0.95, Day of Birth = 0.95, Month of Birth = 0.98, and Year of Birth = 0.90.

Combining these with the u -probabilities derived earlier, the agreement weights A_k , for each of the variables are, 6.8432, 6.2576, 2.3168, 2.4596, and 4.0917 respectively. The disagreement weights D_k of each variable are, -2.3016 , -2.9939 , -2.8974 , -3.8245 , and -2.2874 .

As a result the maximum weight two records can have is 21.9689 i.e. the individuals agree on all variables. The minimum weight two records can have is -14.3049 i.e. the records disagree on all the variables.

A histogram plot of all the weights for each match are provided in Figure 4.2, where all the pairs with weight greater than 20 are declared matches. All pairs with weight less than 5 are declared as unmatched. All pairs with weight between 5 and 20 are manually investigated and decided by a clerical review whether they belong in the matched set or unmatched set.

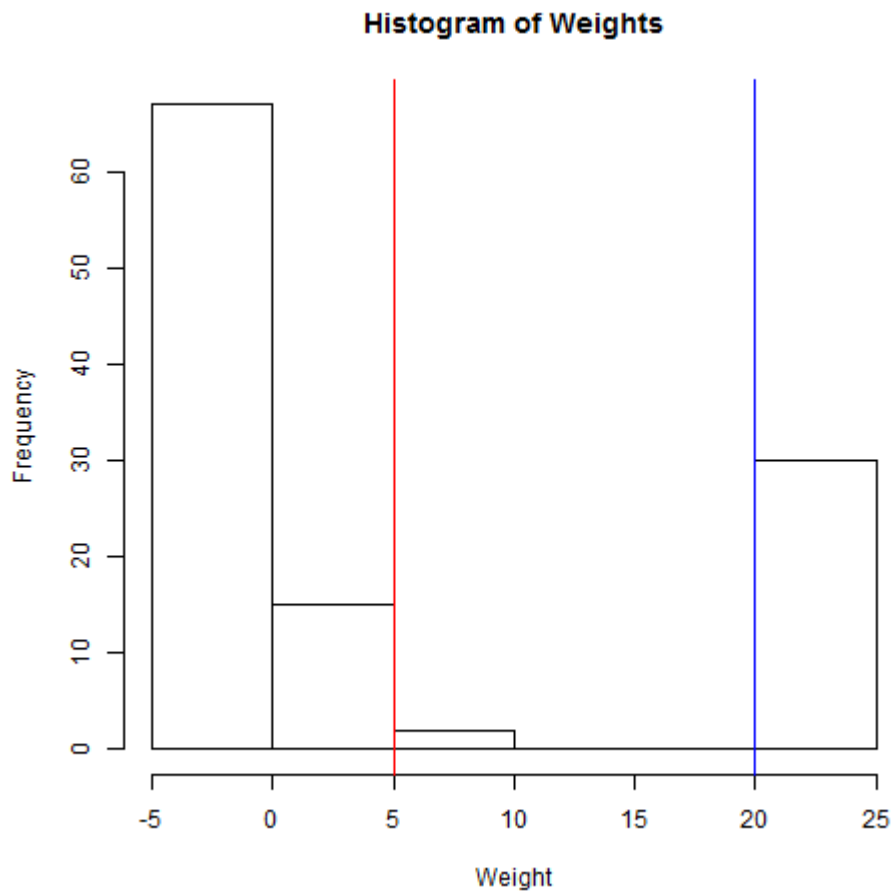


Fig. 4.2: Histogram of Weights used in the Example

4.2.5 Selecting Matching Variables

The next step is to decide which variables are going to be used as identifiers. The reason that all variables are not used as identifiers is weaker matching variables will still contribute to each comparison weight and effect the decision if they are a match or not. In order to decide which variables are effective to match on, the baseline matched dataset is created using all the variables. An algorithm is now designed to determine which variables should be used as the identifiers.

This algorithm identifies all matched pairs using all variables available, then moves through a sequence of steps of adding and removing variables at each stage increasing the number of matched pairs. This is repeated until re-introducing or removing, variables do not increase the number of matched pairs.

Let $\Omega = (v_1, v_2, \dots, v_k)$ be the set of all variables available to match on.

At each iteration t of the algorithm there are always two sets of variables v^t and \bar{v}^t . v^t are the variables remaining at iteration t , in which $v^t \subseteq \Omega$. \bar{v}^t is the set of variables that have been removed in iteration t , so $\bar{v}^t = \Omega \setminus v^t$.

At each iteration t the weights w_{ij}^t are calculated by using the variables in set v^t . Based on the weights w_{ij}^t a match is declared between $a_i \in A$ and $b_j \in B$ if $w_{ij}^t \geq C^t$ where C^t is a decided cutoff weight.

Thus, using the chosen weights, w_{ij}^t , the matching indicator array Φ_{ij}^t is created

$$\Phi_{ij} = \begin{cases} 1 & \text{if } w_{ij}^t \geq C^t \\ 0 & \text{if not} \end{cases} \quad (4.14)$$

where:

$$\Phi_{ij}^t = I(w_{ij}^t \geq C^t),$$

with C^t chosen so that the conditions

$$\sum_{i=1}^{n_A} \Phi_{ij}^t \leq 1 \text{ for all } j,$$

and

$$\sum_{j=1}^{n_B} \Phi_{ij}^t \leq 1 \text{ for all } i,$$

are met. $\nu^t = \sum_{ij} \hat{\Phi}_{ij}^t$ is the estimated number of unique matches in the matched dataset at time t . $\hat{\Phi}_{ij}^t = \hat{\Phi}_{ij}(v^t)$ is the linkage array for the variables in the set v^t at iteration t .

For each variable $k \in \bar{v}^t$ the number of individuals matched when variable k is re-introduced is defined by:

$$\nu_k^t = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \hat{\Phi}_{ij}(\{v^t, k\}) \quad k \in \bar{v}^t. \quad (4.15)$$

In cases when a variable k is removed, the number of individuals matched is defined by:

$$\nu_k^t = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \hat{\Phi}_{ij}(v^t \setminus k) \quad k \in v^t. \quad (4.16)$$

An exhaustive search through all the variables is impossible due to the complexity of the algorithm. So instead a sequence of deletions/re-introductions is constructed to find the best set of variables.

At each stage of the algorithm we have the sets ν and $\bar{\nu}$, to decide which variable is to be removed or re-introduced, ν_k^t is calculated for all variables.

We will only re-introduce variables where the number of matches increases:

$$\nu_k^t > \nu^{t-1} \quad k \in \bar{\nu}^t,$$

and we will only remove variables where the number of matches stays the same or there is an increase:

$$\nu_k^t \geq \nu^{t-1} \quad k \in \nu^t.$$

Recall the weights for each comparison is calculated by:

$$w_{ij}^t = \sum_{k \in v^t} [\gamma_{ijk} A_k + (1 - \gamma_{ijk}) D_k],$$

this simplifies to

$$w_{ij}^t = \sum_{k \in v^t} q_{ijk} \quad (4.17)$$

where:

$$q_{ijk} = \gamma_{ijk} A_k + (1 - \gamma_{ijk}) D_k.$$

Using the matched dataset Φ_{ij}^t the number of individuals that match for each variable can be calculated. In order to do this $\Phi_{ij}^t \gamma_{ijk}$ must be calculated. This is the comparison matrix for each pair of individuals ij and each variable k matched individual, from this the number of individuals that agree on variable k at iteration t is calculated by:

$$\Omega_k^t = \sum_{i=1}^{N1} \sum_{j=1}^{N2} \Phi_{ij}^t \gamma_{ijk}. \quad (4.18)$$

Using this information, how much each variable influences the weight at iteration t can be calculated and is denoted as Ψ_k^t . This is calculated by:

$$\begin{aligned} \Psi_k^t &= \sum_{ij} (\gamma_{ijk} \hat{\Phi}_{ij} \log(\frac{m_k}{u_k}) + (1 - \gamma_{ijk}) \hat{\Phi}_{ijk} \log(\frac{1 - m_k}{1 - u_k})) \\ &= \Omega_k^t A_k + (\nu^t - \Omega_k^t) D_k \end{aligned}$$

where ν^t is the number of individuals that have been declared a match.

When adding a variable Ψ_k^t becomes:

$$\Psi_k^t = \sum_{ij} q_{ijk} \hat{\Phi}_{ij}(\{v^t, k\}) \quad k \in \bar{v}^t, \quad (4.19)$$

and when removing a variable Ψ_k^t is:

$$\Psi_k^t = \sum_{ij} q_{ijk} \hat{\Phi}_{ij}(\{v^t \setminus k\}) \quad k \in v^t, \quad (4.20)$$

The procedure to define what variables are to be used as the identifiers is defined as:

1. Calculate the number of individuals matched using the all variables as identifiers.
2. Calculate the number of individuals matched for each variable when variable is removed from v^t or re-introduced from \bar{v}^t .
3. Look at the Ψ_k^t for all variables that has a maximum increase in the number of individuals matched.
4. Re-introduce/remove the variable with the lowest Ψ_k^t from this subset of variables.

5. Repeat until removing or re-introducing variables do not increase the number of matched individuals.

Re-introduction of variables will only occur if there is an increase in the number of individuals matched. Variables are removed only if there is an increase or no change in the number of individuals matched.

The main idea of this algorithm is it travels through a path always increasing in the number of individuals matched. This approach is continually taken until removing variables provides fewer unique matches. The re-introduction side is only used to check that a stronger combination of variables isn't missed. An effective identifier may be removed in the process the re-introduction step is a check for this occurrence.

Using this algorithm along with all the previous parts of this section we establish the set of matched individuals. Until this point no discussion of the quality of these matches has been discussed. This is covered in Section 4.2.6.

4.2.6 Quality of Matches

Once we have our matching variables and the comparison weights w_{ij} have been established, the next phase is to determine the quality of match, which individuals that are definite matches have a high quality, and individuals that are not as likely a true match have a low quality.

In order to compute the quality of each match we use the algorithm:

Starting with step $l = 1$

1. Identify weight cut off C_l as the lowest weight for which all comparisons above C_l are unique.
2. Identify the unique pairs and label this set of matches as Q_l
3. For individuals $i \in A$ in Q_l set comparison weight w_{ij} to $\min(w_{ij})$
4. For individuals $j \in B$ in Q_l set comparison weight w_{ij} to $\min(w_{ij})$
5. Set l to $l + 1$ and return to step 1 until Q_l is an empty set.

From this algorithm the set of matched pairs are grouped into, $Q = (Q_1, Q_2, Q_3, \dots, Q_l)$ where, as l increases the quality of the matches decreases. Most of the comparisons that are actual matches are likely to be contained in Q_1 , and Q_2 , comparisons that are less likely matches, are contained in Q_3 to Q_l .

This is a slightly different method than having a pre-determined cutoff value before investigating the potential matches. In the traditional approach weights above the upper bound cutoff are considered matches, while weights below the lower bound cutoff are considered unmatched. Weights that fall in between these two values require clerical review to determine if two individuals are considered a match.

Using the quality of the matches, we establish the matched individuals as individuals with the best quality. The clerical review is now replaced with checking the matched quality as the quality decreases.

This establishes the theory this study uses to determine a matched dataset. This theory is applied to the 2004 and 2008 Narcotics Anonymous surveys in Chapter 5.

5. MATCHING THE TWO NA SURVEYS

As we are determining the effectiveness of the Narcotics Anonymous programme; and the ability of individuals to stay clean (drug free) is the key point of interest, investigation around individuals persisting in the population is the best approach. The analysis to determine which individuals are matched, is conducted by separating the dataset into three different groups.

- Individuals with Clean Time Longer than 10 years (prior to 1995);
- Individuals with Clean Time Longer than 5 years (prior to 2000);
- All individuals.

The first part of the analysis is conducted with individuals who have a clean time longer than 10 years. Individuals who have stuck with the programme for more than 10 years (at the start of the study), should remain and be a reliable source of information for the rest of the study. The results from this stable sub-population establishes base estimates for the slack and m -probabilities to use for the second part of the analysis. Individuals with clean time of greater than 5 years are then analysed to strengthen any assumptions made by analysing individuals with clean time longer than 10 years.

The matched individuals from the clean time longer than 10 years are expected to remain matched in clean time longer than 5 years. The final analysis conducted on all individuals is used to determine the actual matched individuals across the 2004 and 2008 surveys. This approach is done under the assumption, individuals with a clean time longer than 10 years are more likely to have stable characteristics and behaviour than individuals just becoming clean.

5.1 Grouping the variables

The analysis is started, by dividing the set of variables V into three separate distinct groups, Group A, Group B, and Group C.

- Group A is defined as the group in which the variables are time invariant and thus are most likely to get a strong match.
- Group B is defined as the group of variables that can change over time, but this change is expected to be small. This group of variables are less likely to get a strong match, but they will still make a match.
- The final group; Group C is the variables which are highly time variant and are expected to change over time. Due to this time variant quality, we expect the variables in Group C are not as helpful in identifying a true match.

The variables *First NA Date* and *Clean Date* have been broken up into three separate parts, Day, Month, and Year. This is because an individual may find it easier to recall the first year they joined NA but may struggle to recall the exact day they joined. Clean date is considered almost as a second birthday to an individual, and is more likely an individual will be able remember the exact day they became clean.

The variables that are included in Group A are:

- Age;
- Sex;
- Ethnicity;
- PreNAIncome;
- PreEducation;
- PreMedical;
- PreMental;
- FirstNAYear;
- CleanYear;
- CleanMonth.

These variables have been grouped into Group A as all are time invariant variables. *Ethnicity* is self identified, and so may change over time, however,

it is reasonable to expect it will be stable for most respondents. *Age* will change over time but this change is known to be the four year period, which the *Age* variable can be easily adjusted to include. Sex is another variable we do not expect to change over time.

PreEducation, *Pre NA income Source*, *PreMedical*, and *PreMental* are time invariant as these are all set before the individual has joined NA. Since the study is interested in the effectiveness of the NA programme, having these variables as time invariant is a crucial way of accessing its effectiveness.

The variable of *PreNAWork* has not been put into group A due to being another self identified variable, where the respondent could consider their actual job was best suited to another category from their response in the survey, to their response in the second survey.

The respondents were asked for their clean date and first NA date in the survey, using this information and the date the individual completed the survey we computed the clean time and NA time. The clean date may match on year and month but fail to match on the day, as individuals may struggle to remember the exact day of certain events. For this reason *FirstNADay*, and *FirstNAMonth* have been excluded from group A as well.

The clean date is to be considered a time invariant variable under this subset only, as individuals with clean time greater than 10 years in 2004 are highly likely to stay clean over the four year period of the study and have greater than 14 years clean time in 2008. Under normal circumstances this variable is highly time dependent, this is due to individuals joining the programme, becoming clean, then relapsing and becoming clean again causing their clean date to change continuously.

The variables that are grouped into Group B are:

- Clean Day;
- First NA Month;
- First NA Day;
- Clean Location;
- Drug Time;
- Drug Choice;
- Most Drug;
- Criminal;

- City;
- Region;
- Most Influence.

CleanDay, *FirstNAMonth*, and *FirstNADay* have been classified into Group B due to the weaker match for the time invariant aspect. *CleanLocation*, *DrugTime*, *DrugChoice*, and *MostDrug* are Group B variables because of the personal interpretation of the individual at time of the survey. This interpretation comes down to the individuals thoughts about their drug use around the time of the survey.

City and *Region* have been grouped into Group B as both are variables we expect to change over time. However this change might only be a subtle change. E.g. an individual moving to another suburb in the same city.

Most Influence is a self identified variable by the respondent, and the way they identify what influenced them the most to join NA may change, as there may be multiple reasons an individual states to why they started attending NA. This means that an individual may state one reason in 2004, and a different reason to why they joined in 2008.

The variables grouped into Group C are:

- StartSteps;
- AllSteps;
- IncomeSource;
- Education;
- PaidWork;
- OngoingMedical;
- OngoingMental;
- PostMedical;
- PostMental;
- Sponsor;
- SponsorOthers;
- OtherFellowships;
- OtherSupport.

These variables are highly time dependent as an individual would be expected to improve as they age and as they work through the programme.

These improvements would mean that the individual may gain higher qualifications, make career changes, and have improved health aspects both mentally and medically. The *Start Steps* variable is one of the most time dependent variables. It refers to whether or not a person has started working through the 12 Steps of the NA Recovery Programme. This is due to the fact that even though an individual might not have started the Steps in 2004 they may easily have started them in 2008. A similar argument may be made for *AllSteps*.

The variables, *Sponsor*, *SponsorOthers*, *OtherFellowships*, and *OtherSupport* have all been grouped into Group C. These variables have been grouped into Group C due to the high time dependence nature of each variable. Individuals are more likely to require a sponsor at the start of their recovery, but decide they no longer require a sponsor after time in the programme. This argument also explains why the other variables are in Group C.

5.2 Clean Time Prior to 1995

We now carry out our matching algorithm on the subset of the data, where respondents had clean date before January 1 1995. We give full details of the procedure and determine parameters for our algorithm. We will apply to the set with clean time prior to January 1 2000 (Section 5.3), and the whole dataset (Section 5.4).

5.2.1 Determining u -probability

Recall the definition of u -probability from Section 4.2.2

$$u_k = \Pr(a_i \text{ and } b_j \text{ agree on variable } k \mid a_i \text{ and } b_j \text{ are not a true match}),$$

which can be estimated by:

$$u_k = \frac{\sum_i \sum_j \gamma_{ijk}}{n_A n_B}, \quad (5.1)$$

where n_A is the number of individuals who complete the 2004 survey with a clean time prior to 1995 and n_B is the number of individuals who complete the 2008 survey with clean time prior to 1995. Here $n_A = 75$, and $n_B = 85$. As a result $n_A n_B = 75 \times 85 = 5950$.

The slack δ_k that is used in calculating these u -probabilities is derived in Section 5.2.3.

Using this information each of the u -probabilities can be calculated. To calculate the u -probabilities now all that is required is the number of “agreements” on each variable. These agreements are the number of times individuals in 2004 agreed with individuals in 2008 on selected variable. For *FirstNADay* there are 5280 agreements giving the u -probability of $\frac{5280}{5950} = 0.8874$, all the u -probabilities are given in the Table 5.1.

Tab. 5.1: Table of u -probabilities

Variable	# Agreed	$n_A n_B$	u -probability
FirstNADay	5280	5950	0.8874
FirstNAMonth	2026	5950	0.3405
FirstNAYear	1285	5950	0.216
Sex	3185	5950	0.5353
Age	686	5950	0.1153
City	1057	5950	0.1776
Ethnicity	3074	5950	0.5166
Sponsor	4236	5950	0.7119
SponsorOthers	3641	5950	0.6119
CleanLocation	1667	5950	0.2802
StartSteps	5646	5950	0.9489
AllSteps	5311	5950	0.8926
OtherFellowship	3394	5950	0.5704
OtherSupport	4430	5950	0.7445
MostInfluence	1656	5950	0.2783
DrugTime	2497	5950	0.4197
DrugChoice	799	5950	0.1343
MostDrug	1238	5950	0.2081
IncomeSource	1881	5950	0.3161
PreNAIncome	1544	5950	0.2595
Education	1327	5950	0.223
PreNAEducation	1515	5950	0.2546
PaidWork	1121	5950	0.1884
PreNAWork	729	5950	0.1225
PreMedical	2618	5950	0.44
PreMental	2563	5950	0.4308
OngoingMedical	3600	5950	0.605
OngoingMental	4025	5950	0.6765
PostMedical	2240	5950	0.3765
PostMental	2986	5950	0.5018
Criminal	2899	5950	0.4872
Region	1958	5950	0.3291
CleanDay	2427	5950	0.4079
CleanMonth	1453	5950	0.2442
CleanYear	623	5950	0.1047

5.2.2 Estimating m -probability

Baseline estimates for the m -probabilities for each variable are derived by the using the grouping of the variables. Starting off with all the variables in group A, the assumption is made that group A are the most reliable. Using this assumption a second assumption is made that most of these variables match 95% -98% of the time. When the group is broken down further the variables that are assumed that will always match are, *Sex*, *Ethnicity*, and *CleanYear*. However due to errors that maybe encountered throughout the process these three variables should have a m -probability of 0.98, with all other variables in group A having m -probabilities of 0.95. This is because these variables are likely to be a true match because people will know confidently their sex, ethnicity, and year they became clean but may become a bit unclear on other aspects.

Next the m -probabilities for each variable in group B are estimated. These variables are considered less reliable than all variables in group A. This group is also broken down further, which is done by making assumptions about how well individuals in the study remember details between the survey period.

The variables *CleanLocation*, *MostDrug*, *DrugTime*, *Criminal*, and *Region* are all variables which the individual is assumed to remember due to the nature around the individuals becoming clean. *MostDrug* is assumed to be information which the individual remembers. However, this variable has been placed in group B as individuals may take two or more drugs equally, or their opinion about their drug use may change between the survey. Because of this logic we assign the m -probability of 0.95. A similar argument is made for *CleanLocation*, *DrugTime*, *MostInfluence*, and *DrugChoice*. These get assigned the m -probabilities of 0.95, 0.92, 0.9, 0.9 respectively. This is due to if an individual truly remembers where they became clean they will, they may be a bit more unclear on how long they used drugs but will roughly know how long, and they may have a different opinion on what their favourite drug was or what caused them to become clean between 2004 and 2008.

The variables *Criminal*, *Region* and *City* do not follow this same logic, as individuals are likely to be matched on whether they have criminal records based on their drug use. Assuming the individuals stay clean during the four year period between surveys, the assumption, that the status about the *Criminal* variable should also remain the same is made. However, due to the sensitive nature of the variable more individuals are less likely to answer. For this reasons it gets assigned the base m -probability of 0.95. Individuals

are assumed to remain in the same city during the four year survey period. Due to the nature of individuals changing cities for various reason the *City* variable is assigned an m -probability of 0.85. In cases where individuals shift city, the assumption they will remain in the same region is made. Because of this assumption the base m -probability for *Region* is 0.95.

The remaining variables *FirstNADay*, *FirstNAMonth*, and *CleanDay* are assumed to be remembered by the individuals, using the same argument from *CleanMonth* and *CleanYear*. The assumption that individuals are less likely to remember the exact day is made. But, the individual should remember the rough time they joined NA and the day they became clean. The variables *FirstNADay* and *FirstNAMonth* have the respective m -probabilities of 0.9 and 0.93. These m -probabilities have been assigned as an individual may remember the exact moth they joined and will be a true match, but they may find it harder to remember exact day reducing the chance of a true match. For *CleanDay* an individual may recall the exact week when they became clean however they may redefine the exact criteria of becoming clean, for this reason the m -probability of 0.9 is assigned.

Finally the variables in group C are investigated. The variables *OngoingMedical*, *OngoingMental*, *PostMedical*, and *PostMental* are assumed to be quite volatile between the two survey, as individuals are likely to develop health concerns after stopping drug use. For this reason these variables are given the m -probability of 0.8.

The variables *Sponsor*, *SponsorOthers*, *OtherFellowship*, *OtherSupport*, *IncomeSource*, *Education*, and *PaidWork* are highly dependant on the individual answering as an individual may freely change the status of these variables between the two surveys meaning an individual may have a sponsor in 2004, but circumstances may change and they may no longer have a sponsor by 2008. For this reason, the m -probability of these variables is assigned 0.9.

A similar argument can be used for the *PaidWork* variable. However, the grouping of responses was different from 2004 and 2008 meaning the response from one survey year to another is less reliable. For this reason a lower m -probability of 0.85 is used.

The final set of variables that are discussed are *StartSteps* and *AllSteps*. These variables have the property that if an individual has started the steps or has completed all the steps in 2004 then they should have started or completed them still in 2008. However, conversely if an individual hasn't started the steps in 2004, this doesn't mean they wouldn't have started the steps by 2008. In fact an individual may not have started the steps in 2004,

but started and completed them by 2008. Because of this property we give these variables a m -probability of 0.87.

Now the baseline estimated m -probabilities have been established. The amount of slack that each variable should have can be determined.

5.2.3 Determining Slack

Using the baseline estimates for m -probability the amount of slack that is relevant for each variable is determined. Starting off the type of variable is determined. There are 28 categorical variables, these variables all have a slack of 0, there are 7 numeric variables which have slack element that needs to be determined. These variables are:

- FirstNADay;
- FirstNAMonth;
- FirstNAYear;
- Age;
- CleanDay;
- CleanMonth;
- CleanYear.

There are 3 cases which are considered, these having slack values as shown in Table 5.2. All the weights have been calculated by using the u -probabilities

Tab. 5.2: Table of different slack cases

Variable	Case 1	Case 2	Case 3
FirstNADay	0	3	31
FirstNAMonth	0	1	2
FirstNAYear	0	0	1
Age	0	2	4
CleanDay	0	3	7
CleanMonth	0	1	2
CleanYear	0	0	1

associated with each case of slack in Table 5.2, and our initial estimates of m -probabilities Section 5.2.2 are used to calculate the weights.

Starting with the first case where all variables have no slack allowance, we get the plot for the weights shown in Figure 5.1.

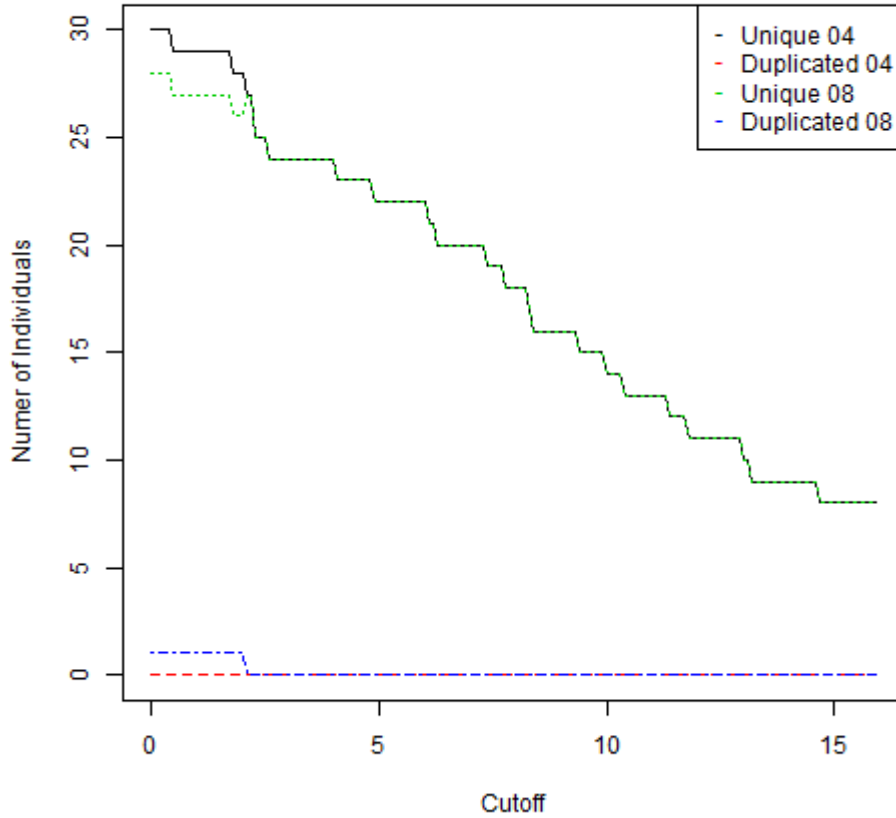


Fig. 5.1: Sensitivity Analysis Case 1: no slack allowed

Using these weights we determine the cutoff weight by scanning through the set of possible values and seeing how the number of matched pairs changes. The number of unique matches are determined by scanning through the weights in decreasing order (right to left). We take the number of individuals matched until we come across the first individual that is matched twice, as the cutoff point.

There are 27 unique matches with a cutoff of 2.1 without allowing any slack (Figure 5.1). Using this as a base case the other two cases are investigated.

Investigating Case 2 by using the slack reported in Table 5.2, the new cutoff weight is determined. As in Case 1 this is achieved by scanning through determining the number of unique matches and taking the left most point.

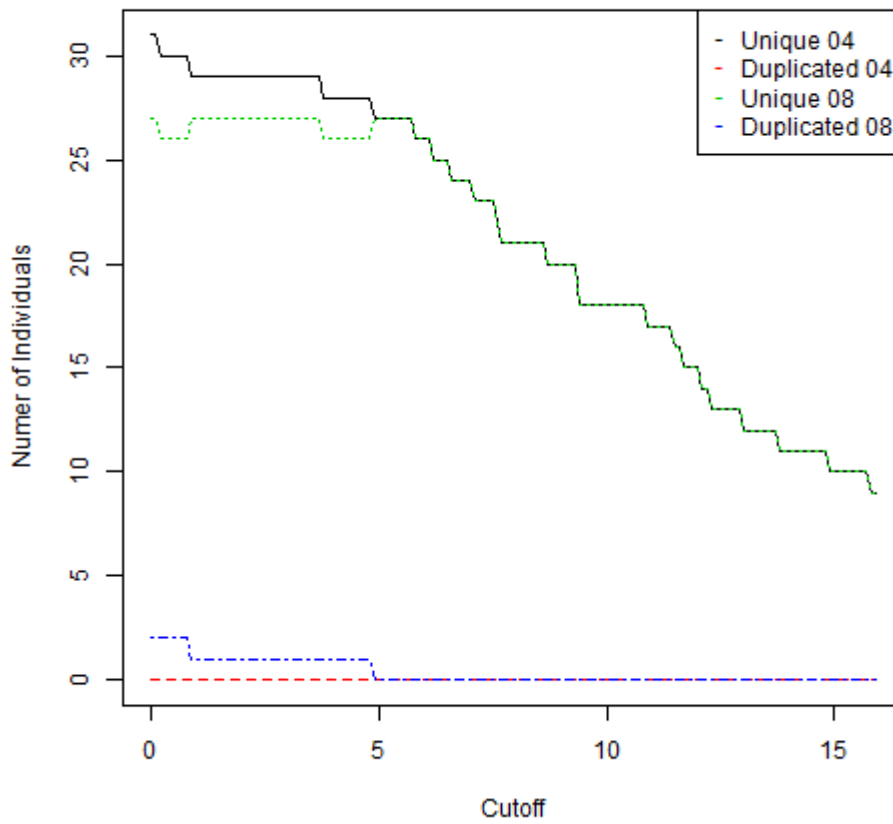


Fig. 5.2: Sensitivity Analysis Case 2: moderate slack

Using the slack values for case 2 there are 27 unique matches again, this time there is a higher cutoff of 4.9 (Figure 5.2). There is no increase or decrease in the number of unique matches, but there is a higher cutoff for the same number of matches. Due to this increase in the cutoff weight this case of an allowance of slack is better than having no slack at all.

Next Case 3 where we have an increased amount of slack for each numeric variable is investigated, case 3 is to be considered as the “high slack” case.

The high slack case has a slack of \pm a week (7 days) for *CleanDay*. For *FirstNADay* a slack of 31 days is introduced, this slack of 31 ensures there will always be a match on this variable. This has been done under the assumption that the individuals with clean time earlier than 1995 would be uncertain on the exact day they had their first meeting but rather more certain about the day they became clean.

A slack of \pm 2 months is assigned to both *FirstNAMonth* and *CleanMonth*, and a slack of \pm 1 year is assigned for both *CleanYear* and *FirstNAYear*. Using these slacks the cutoff is determined as before.

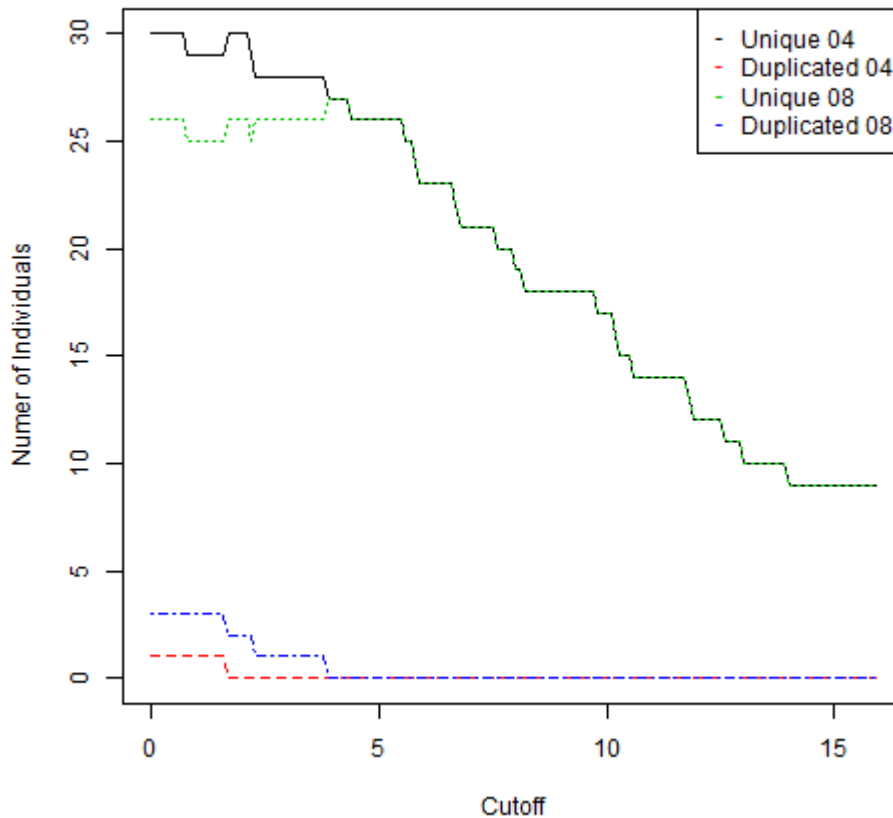


Fig. 5.3: Sensitivity Analysis Case 3: high slack

Again there are 27 unique matches with a cutoff of 3.9 (Figure 5.3) but a lower cutoff than the Case 2 the moderate slack case, which is considered a worse match.

Because the slack from the high slack case brought about a worse match than the moderate slack, we reduce this high slack down closer to the moderate slack and determine whether there is a more ideal case in between the two. This has been achieved by reducing Age variable slack from ± 4 years to ± 1 year.

Using this adjusted slack we determine the cutoff by scanning through determining the number of unique matches for both 2004 and 2008 data.

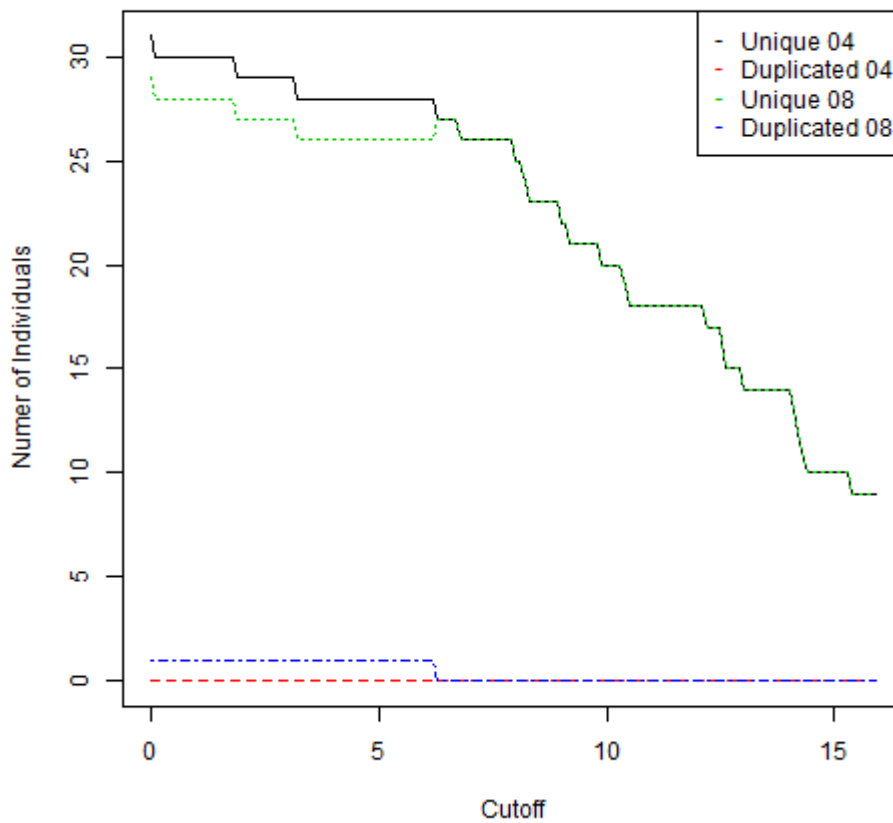


Fig. 5.4: Sensitivity Analysis of chosen slack

There are 27 unique matches again with a cutoff of 6.3 (Figure 5.4). Since this case has the highest cutoff any reduction would tend towards the moderate slack case. This is the slack that is used for our investigation, and is presented in Table 5.3.

Tab. 5.3: Table of chosen slack

Variable	Slack
FirstNADay	31
FirstNAMonth	2
FirstNAYear	1
Age	1
CleanDay	7
CleanMonth	1
CleanYear	0

5.2.4 Determining m -probability

Using the slack, the associated u -probability and the baseline estimates for the m -probabilities determined in the previous sections, we determine the weights and set of matched individuals, and matched set. From this match set we calculate the empirical m -probabilities and re-adjust our estimates based on these empirical values.

Our initial estimates for the u -probabilities, m -probabilities, and slack are given in Table 5.4.

From these initial estimates there are 27 matches. Using these 27 matches the empirical m -probabilities are calculated in Table 5.4.

Tab. 5.4: Table of Empirical m-probability

Variable	Slack	m-prob	u-prob	# Matched	Emp m-prob
FirstNADay	31	0.9	0.8874	27	1
FirstNAMonth	2	0.93	0.3405	19	0.7037
FirstNAYear	1	0.95	0.216	25	0.9259
Sex	0	0.98	0.5353	27	1
Age	1	0.95	0.1153	26	0.963
City	0	0.85	0.1776	25	0.9259
Ethnicity	0	0.98	0.5166	25	0.9259
Sponsor	0	0.9	0.7119	23	0.8519
SponsorOthers	0	0.9	0.6119	24	0.8889
CleanLocation	0	0.95	0.2802	19	0.7037
StartSteps	0	0.87	0.9489	26	0.963
AllSteps	0	0.87	0.8926	26	0.963
OtherFellowship	0	0.9	0.5704	23	0.8519
OtherSupport	0	0.9	0.7445	24	0.8889
MostInfluence	0	0.9	0.2783	19	0.7037
DrugTime	0	0.92	0.4197	23	0.8519
DrugChoice	0	0.9	0.1343	20	0.7407
MostDrug	0	0.96	0.2081	21	0.7778
IncomeSource	0	0.9	0.3161	20	0.7407
PreNAIncome	0	0.95	0.2595	18	0.6667
Education	0	0.9	0.223	21	0.7778
PreNAEducation	0	0.95	0.2546	22	0.8148
PaidWork	0	0.85	0.1884	19	0.7037
PreNAWork	0	0.95	0.1225	14	0.5185
PreMedical	0	0.95	0.44	21	0.7778
PreMental	0	0.95	0.4308	18	0.6667
OngoingMedical	0	0.8	0.605	21	0.7778
OngoingMental	0	0.8	0.6765	25	0.9259
PostMedical	0	0.8	0.3765	16	0.5926
PostMental	0	0.8	0.5018	20	0.7407
Criminal	0	0.95	0.4872	25	0.9259
Region	0	0.95	0.3291	26	0.963
CleanDay	7	0.9	0.4079	25	0.9259
CleanMonth	1	0.95	0.2442	26	0.963
CleanYear	0	0.98	0.1047	26	0.963

Since the empirical m -probabilities are an upper limit for the m -probabilities, the estimates of our original m -probabilities are re-adjusted accordingly.

For all the estimates that have been over estimated we drop our estimates down below the empirical probabilities. For the estimates that have been under-estimated, we adjust our estimates by either bringing closer to the empirical probabilities, or leaving the estimate unadjusted.

The number of unique matches are recalculated by using these new empirical based m -probabilities. Using these new estimated m -probabilities there are 28 matches and because of this increase in the number of unique matches we re-adjust our m -probabilities again for any over/under estimation.

From these re-adjustments only 8 variables have been over estimated, these variables are:

- CleanLocation;
- MostDrug;
- IncomeSource;
- PreNAEducation;
- PaidWork;
- PreMedical;
- PreMental;
- PostMental.

The m -probabilities are re-adjusted respectively 0.67, 0.75, 0.71, 0.78, 0.67, 0.75, 0.64, and 0.72. These new estimated m -probabilities are used to determine the number of unique matches, and the new empirical m -probabilities.

Using these m -probabilities there are 28 unique matches with the same empirical probabilities as before. As we have the same empirical probabilities we use these m -probabilities for a base of the matching.

The concluding table (Table 5.5) of the slack, u -probabilities, and m -probabilities that are used in the analysis is provided.

Tab. 5.5: Table of chosen values

Variable	Slack	m-probability	u-probability
FirstNADay	31	0.98	0.8874
FirstNAMonth	2	0.7	0.3405
FirstNAYear	1	0.92	0.216
Sex	0	0.98	0.5353
Age	1	0.95	0.1153
City	0	0.85	0.1776
Ethnicity	0	0.92	0.5166
Sponsor	0	0.85	0.7119
SponsorOthers	0	0.88	0.6119
CleanLocation	0	0.67	0.2802
StartSteps	0	0.96	0.9489
AllSteps	0	0.9	0.8926
OtherFellowship	0	0.85	0.5704
OtherSupport	0	0.88	0.7445
MostInfluence	0	0.7	0.2783
DrugTime	0	0.85	0.4197
DrugChoice	0	0.74	0.1343
MostDrug	0	0.75	0.2081
IncomeSource	0	0.71	0.3161
PreNAIncome	0	0.66	0.2595
Education	0	0.77	0.223
PreNAEducation	0	0.78	0.2546
PaidWork	0	0.67	0.1884
PreNAWork	0	0.5	0.1225
PreMedical	0	0.75	0.44
PreMental	0	0.64	0.4308
OngoingMedical	0	0.77	0.605
OngoingMental	0	0.85	0.6765
PostMedical	0	0.59	0.3765
PostMental	0	0.71	0.5018
Criminal	0	0.92	0.4872
Region	0	0.95	0.3291
CleanDay	7	0.9	0.4079
CleanMonth	1	0.96	0.2442
CleanYear	0	0.96	0.1047

Next we must determine which variables will be used to base the matching on.

5.2.5 Selecting Match Variables

Following the algorithm:

1. Calculate the number of individuals matched using the all variables as identifiers.
2. Calculate the number of individuals matched for each variable when variable is removed from v^t or re-introduced from \bar{v}^t .
3. Look at the Ψ_k^t for all variables that has a maximum increase in the number of individuals matched.
4. Re-introduce/remove the variable with the lowest Ψ_k^t from this subset of variables.
5. Repeat until no improvement can be found.

The variables that should be used are derived. Starting with all variables being used the initial step of the algorithm is presented in Table 5.6.

Tab. 5.6: Table of Start of process

Variable	# Matched Removed	Ω	agree	disagree	Ψ
FirstNADay	28	28	0.0993	-1.7281	2.7792
FirstNAMonth	28	20	0.7207	-0.7877	8.1117
FirstNAYear	28	26	1.4491	-2.2824	33.1117
Sex	28	28	0.6047	-3.1457	16.9323
Age	28	27	2.1089	-2.8732	54.0677
City	29	26	1.5657	-1.7016	37.3051
Ethnicity	28	26	0.5771	-1.7988	11.4071
Sponsor	28	24	0.1773	-0.6527	1.6444
SponsorOthers	28	25	0.3634	-1.1738	5.5625
CleanLocation	29	19	0.8718	-0.7799	9.5448
StartSteps	28	27	0.0116	-0.2449	0.0691
AllSteps	28	27	0.0083	-0.0714	0.1515
OtherFellowship	28	24	0.3989	-1.0522	5.3647
OtherSupport	28	25	0.1672	-0.7557	1.9130
MostInfluence	28	20	0.9224	-0.8778	11.4250
DrugTime	25	24	0.7057	-1.3529	11.5251
DrugChoice	22	21	1.7066	-1.2029	27.4181
MostDrug	23	21	1.2821	-1.1529	18.8523
IncomeSource	29	20	0.8092	-0.8579	9.3207
PreNAIncome	29	19	0.9335	-0.7784	10.7308
Education	20	22	1.2392	-1.2174	19.9586
PreNAEducation	28	22	1.1196	-1.2203	17.3094
PaidWork	28	19	1.2687	-0.8999	16.0062
PreNAWork	26	14	1.4065	-0.5625	11.8164
PreMedical	28	21	0.5333	-0.8065	5.5539
PreMental	27	18	0.3958	-0.4581	2.5436
OngoingMedical	28	22	0.2412	-0.5408	2.0607
OngoingMental	28	26	0.2283	-0.7686	4.3988
PostMedical	28	17	0.4492	-0.4192	3.0254
PostMental	28	20	0.3471	-0.5411	2.6123
Criminal	29	26	0.6357	-1.8579	12.8125
Region	28	27	1.0601	-2.5966	26.0261
CleanDay	20	26	0.7914	-1.7785	17.0187
CleanMonth	28	27	1.3689	-2.9389	34.0226
CleanYear	28	27	2.2158	-3.1083	56.7192

When *City*, *CleanLocation*, *IncomeSource*, *PreNAIncome*, and *Criminal* are removed more unique matches are obtained. The Ψ value for each of the variables is calculated, and provided in Table 5.7

Tab. 5.7: Table of Start of process Reduced

Variable	# Matched Removed	Ω	agree	disagree	Ψ
City	29	26	1.5657	-1.7016	37.3051
CleanLocation	29	19	0.8718	-0.7799	9.5448
IncomeSource	29	20	0.8092	-0.8579	9.3207
PreNAIncome	29	19	0.9335	-0.7784	10.7308
Criminal	29	26	0.6357	-1.8579	12.8125

IncomeSource has the lowest Ψ value as a result it is removed with an increase from 28 matches to 29 matches.

In the next iteration there is no increase in the number of unique matches. However, there are a number of variables that when removed keep the same number of unique matches. In this situation we remove the variable with the lowest Ψ ; which is *StartStep*, with $\Psi = 0.08073119$. This variable is removed due to the low influence of the agreement weight $A_k = 0.01162987$.

Next *AllSteps* is removed with $\Psi = 0.1597839$ with 29 unique matches. Next *Sponsor* is removed retaining the 29 unique matches with a $\Psi = 1.821783$. *OtherSupport* is removed next maintaining the 29 unique matches $\Psi = 2.080243$. *OngoingMedical* is removed with the smallest Ψ of $\Psi = 2.301889$. The next variable that gets removed is *FirstNADay* with $\Psi = 2.878445$. All these removal have maintained the same 29 unique matches.

In the next iteration *FirstNAYear* is removed with an increase from 29 unique matches to 31 unique matches with a $\Psi = 34.56081$.

The *Criminal* variable is removed next with an increase to 32 unique matches, it gets removed due to the lowest $\Psi = 12.22599$. The other variable considered to be removed with an increase to 32 unique matches is *Age* with a $\Psi = 60.39451$.

Region is the next variable that is removed with a Ψ of 30.26651, with an increase from 32 unique matches to 33 unique matches.

In the next iteration *PreMental* variable is re-introduced. This is because the maximum increase for removing any variable is still 33 matches as demonstrated in Table 5.8

Tab. 5.8: Table of Removal of variable

Variable	# Matched Removed	Ω	agree	disagree	Ψ
Ethnicity	33	31	0.5771	-1.7988	14.2926
PreNAEducation	33	26	1.1196	-1.2203	20.5676

While when we re-introduce *PreMental* an extra unique match is found.

Ethnicity is removed with an increase from 34 unique matches to 35 matches, followed by remove *Education* which has a Ψ of 17.56764

In the next iteration there is no increase in the number of unique matches when a variable is removed, nor is there an increase in the number of unique matches when the variable is introduced.

When any variable is removed there is at least a decrease from 35 unique matches to 33 matches.

Tab. 5.9: Table of Final Iteration Removed

Variable	# Matched Removed	Ω	agree	disagree	Ψ
MostDrug	33	24	1.2821	-1.1530	18.0866
PaidWork	33	21	1.2687	-0.8999	14.04411
PreMedical	33	26	0.5333	-0.8065	6.6075
PreNAEducation	33	27	1.1196	-1.2203	20.4669
PreNAWork	33	14	1.4065	-0.5625	7.8791

when any variable is re-introducing there is a maximum increase of 0 unique matches.

Tab. 5.10: Table of Final Iteration Re-Introduced

Variable	# Matched Removed	Ω	agree	disagree	Ψ
AllSteps	35	34	0.0083	-0.07139	0.2093
PostMental	35	23	0.3471	-0.5411	1.4890
Sponsor	35	31	0.1773	-0.6527	2.8856
StartSteps	35	33	0.01163	-0.2449	-0.1060

Using this information the matching of individuals with clean time earlier than 1995 is:

Tab. 5.11: Table of Information used clean time prior 1995

Variable	slack	m -prob	u -prob
Sex	0	0.98	0.5353
Age	1	0.95	0.1153
City	0	0.85	0.1776
SponsorOthers	0	0.88	0.6119
CleanLocation	0	0.67	0.2802
DrugTime	0	0.85	0.4197
DrugChoice	0	0.74	0.1343
MostDrug	0	0.75	0.2081
IncomeSource	0	0.71	0.3161
Education	0	0.77	0.223
PreNAEducation	0	0.78	0.2546
PaidWork	0	0.67	0.1884
PreNAWork	0	0.5	0.1225
PreMedical	0	0.75	0.44
PreMental	0	0.64	0.4308
CleanDay	7	0.9	0.4079
CleanMonth	1	0.96	0.2442
CleanYear	0	0.96	0.1047

Once all the potential matches have been made the comparison weights w_{ij} for individual i in 2004 and individual j in 2008 are calculated. In addition we calculate the $\min(w(i, j))$. The quality of match must now be determined. This is done by using the algorithm defined earlier in Chapter 4, starting with $l = 1$

1. Identify weight cut off C_l as the lowest weight for which all comparisons above C_k are unique.
2. Identify the unique pairs and label as Q_l
3. For individuals i in Q_l set comparison weight $w(i, j)$ to $\min(w(i, j))$,
4. For individuals j in Q_l set comparison weight $w(i, j)$ to $\min(w(i, j))$,
5. set l to $l + 1$ and return to step 1 until Q_l is an empty set.

Apply this process a cutoff weight for the highest quality of 5.9323 is obtained, with the next highest quality being set at -0.5242 .

For the 70 possible matches that can occur, we have the following quality distribution for individuals with clean time prior to 1995, provided in Table 5.12.

Tab. 5.12: Table of Quality Distribution prior to 1995

Matched set	Frequency
Q1	35
Q2	11
Q3	1
Q4	4
Q5	2
Q6	9
Q7	3
Q8	2
Q9	1
Q10	2

The representation of this distribution is shown by Figure 5.5

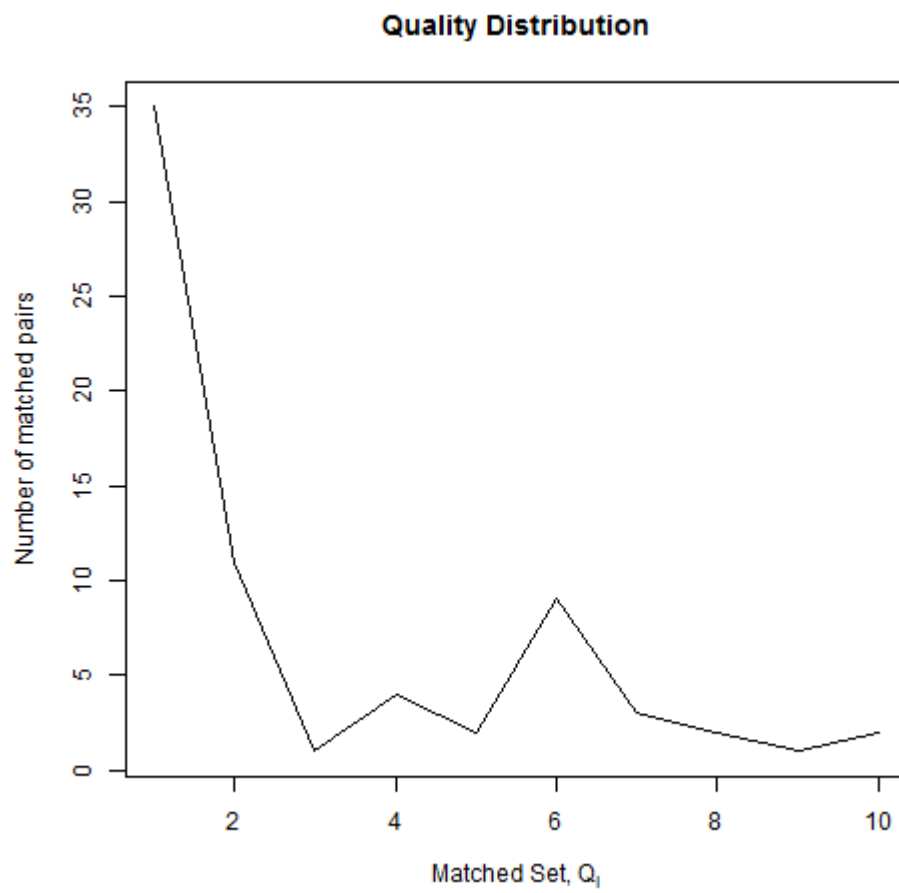


Fig. 5.5: Quality Distribution individuals with clean time earlier than 1995

Using this information as a base, the introduction of individuals with clean time prior to 2000 is now analysed.

5.3 Clean Time Prior to 2000

We now repeat the analysis above for individuals who were clean for more than 5 years in 2004. Because more individuals are introduced, the size of the comparison space has increased, there are now 151 individuals from the 2004 survey and 121 individuals from the 2008 survey. As a result the u -probabilities have all changed, the slack, and m -probabilities have remained the same.

Starting with the information provide in Table 5.13

Tab. 5.13: Table of initial information used clean time prior 2000

Variable	slack	m -prob	u -prob
Sex	0	0.98	0.5235
Age	1	0.95	0.1015
City	0	0.85	0.1755
SponsorOthers	0	0.88	0.5553
CleanLocation	0	0.67	0.2961
DrugTime	0	0.85	0.4129
DrugChoice	0	0.74	0.1241
MostDrug	0	0.75	0.1997
IncomeSource	0	0.71	0.3664
Education	0	0.77	0.2268
PreNAEducation	0	0.78	0.2532
PaidWork	0	0.67	0.1875
PreNAWork	0	0.5	0.1259
PreMedical	0	0.75	0.4275
PreMental	0	0.64	0.4144
CleanDay	7	0.9	0.4069
CleanMonth	1	0.96	0.2427
CleanYear	0	0.96	0.0724

There is an initial matching of 30 individuals, and there is an initial loss of 5 unique matches from the matched dataset of individuals with clean time prior to 1995.

When *DrugTime* is removed there is an increase to 39 matches with $\phi = 42.46893$. In next iteration *Education* is removed with an increase from 39 matches to 48 matches with a $\phi = 36.04663$.

Next *City* is removed, with an increase from 48 unique matches to 51 unique matches where $\phi = 65.35184$. MostDrug is removed after with an increase of 51 unique matches to 54 unique matches with a $\phi = 50.77461$. Next *PreMedical* is removed with an increase to 56 matches with a $\phi = 12.72083$.

Next *PostMental* is reintroduced, this has an increase from 56 unique matches to 59 unique matches.

Using the following information from Table 5.14 as a base to match individuals with clean time prior to 2000 individuals.

Tab. 5.14: Table of Information used clean time prior 2000

Variable	slack	m -prob	u -prob
Sex	0	0.98	0.5235
Age	1	0.95	0.1015
SponsorOthers	0	0.88	0.5553
CleanLocation	0	0.67	0.2961
DrugTime	0	0.85	0.4129
IncomeSource	0	0.71	0.3664
PreNAEducation	0	0.78	0.2532
PaidWork	0	0.67	0.1875
PreNAWork	0	0.5	0.1259
PreMental	0	0.64	0.4144
PostMental	0	0.71	0.56
CleanDay	7	0.9	0.4069
CleanMonth	1	0.96	0.2427
CleanYear	0	0.96	0.0724

Going through the quality check for the data there are 9 different levels of quality until the algorithm finishes. Individuals with quality 1-4 are declared matched. Using this matching decision, there are 83 matches out of a possible 121 individuals matched.

All 35 individuals that were matched with clean time prior 1995, have also been matched in this set of 83 individuals matched. From these 83 individuals 5 additional individuals with a clean time prior to 1995 have now been matched. The quality distribution is given in Table 5.15.

Tab. 5.15: Table of Quality Distribution prior to 2000

Matched Set	Frequency
Q1	59
Q2	11
Q3	3
Q4	10
Q5	1
Q6	2
Q7	6
Q8	8
Q9	1

This distribution is represented by Figure 5.6

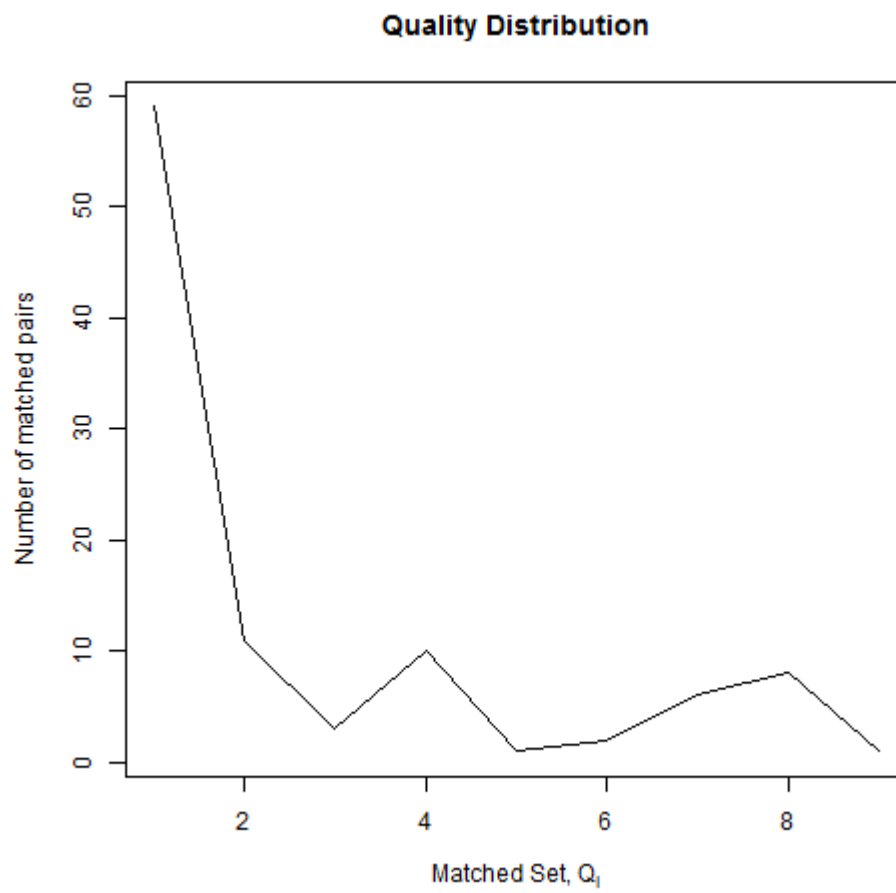


Fig. 5.6: Quality Distribution individuals with clean time earlier than 2000

We now move on to analysing the full dataset.

5.4 Complete Dataset

Finally, we create our matched dataset for use in the analysis of population dynamics for individuals in NA, using all the available records. Again, the size of the comparison space has increased, there are now 475 individuals from the 2004 survey and 546 individuals from the 2008 survey. As a result the u -probabilities change for all variables, and the slack and m -probabilities have remained the same.

Using the final resulting information from individuals with clean time prior to 2000, as a starting base (Table 5.16):

Tab. 5.16: Table of initial information used complete dataset

Variable	slack	m -prob	u -prob
Sex	0	0.98	0.5091
Age	1	0.95	0.0787
SponsorOthers	0	0.88	0.5880
CleanLocation	0	0.67	0.2944
DrugTime	0	0.85	0.4072
IncomeSource	0	0.71	0.2696
PreNAEducation	0	0.78	0.2185
PaidWork	0	0.67	0.1448
PreNAWork	0	0.5	0.1153
PreMental	0	0.64	0.3875
PostMental	0	0.71	0.5699
CleanDay	7	0.9	0.3507
CleanMonth	1	0.96	0.2241
CleanYear	0	0.96	0.02779

We have an initial match of 17 individuals. When *MostDrug* is reintroduced there is an increase to 26 individuals. Next *Education* is reintroduced with an increase from 26 unique matches to 40. *OngoingMental* is reintroduced with an increase to 51 unique matches.

Next *IncomeSource* is removed with an increase from 51 unique matches to 60 unique matches with a $\Psi = 31.2667$. *CleanLocation* is removed next with an increase to 72 unique matches with $\Psi = 22.5766$. The next variable that gets removed is *CleanDay* with a $\Psi = 51.83908$ which has 79 unique matches.

Finally, *PostMedical* is re-introduced with 82 unique matches. As a result the information in Table 5.17 is used as the matching base for all individuals:

Tab. 5.17: Table of information used for complete dataset

Variable	slack	m -prob	u -prob
Sex	0	0.98	0.5091
Age	1	0.95	0.0787
SponsorOthers	0	0.88	0.59
DrugTime	0	0.85	0.4072
MostDrug	0	0.75	0.1692
Education	0	0.77	0.1945
PreNAEducation	0	0.78	0.2185
PaidWork	0	0.67	0.1448
PreNAWork	0	0.5	0.1153
PreMental	0	0.64	0.3875
OngoingMental	0	0.85	0.5796
PostMedical	0	0.59	0.5343
PostMental	0	0.71	0.5699
CleanMonth	1	0.96	0.2241
CleanYear	0	0.96	0.0278

Using the information given in Table 5.17, the quality distribution for all individuals that have been matched is now given in Table 5.18.

Tab. 5.18: Table of Quality Distribution

Matched Set	Frequency
Q1	82
Q2	27
Q3	15
Q4	13
Q5	2

This distribution is represented by Figure 5.7.

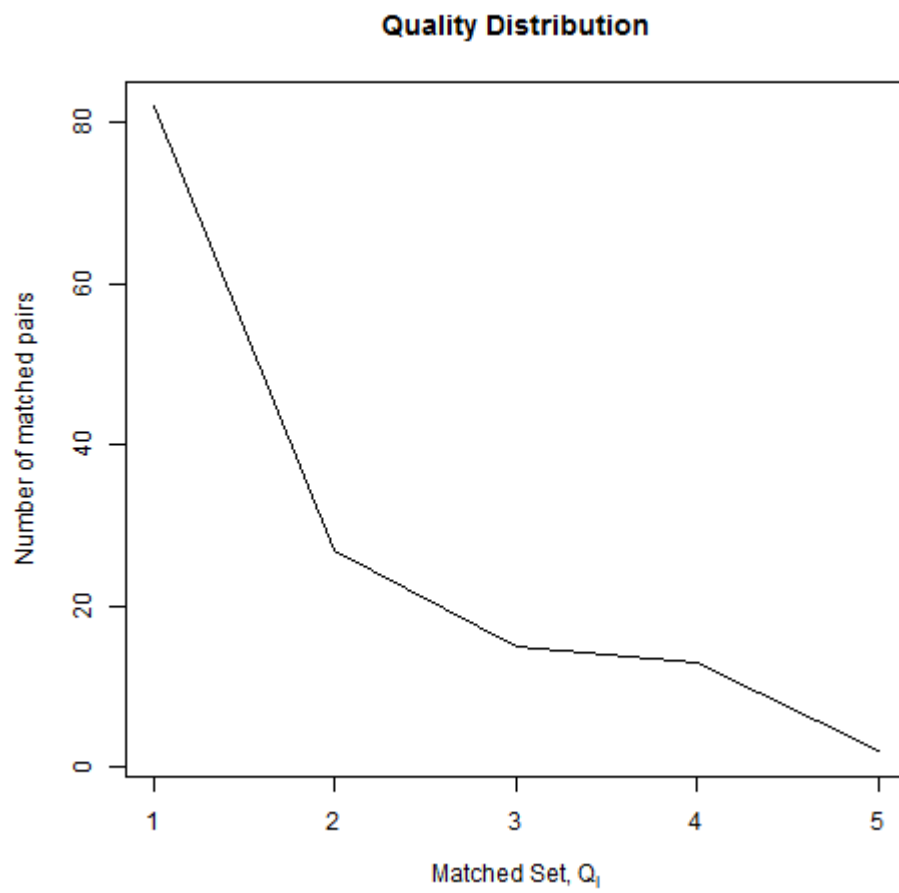


Fig. 5.7: Quality Distribution

As a result, 139 individuals from 2004 have been matched in 2008, with 336 declared as non-matches.

After investigating these records there are 2 matches no longer present for individuals with clean time prior to 1995, and 18 matches no longer present for the individuals with clean time prior to 2000.

The summary information of the probabilistic matching is shown in Table 5.19, where for the full data set we have a 29% match rate. This match rate is calculated by the number of matches divided by the number of total possible matches (if everyone in the smaller set is uniquely matched to another individual in the larger set).

Tab. 5.19: Summary Table of matches

	Clean Time Prior 1995	Clean Time Prior 2000	Full Dataset
# in 2004	70	151	475
# in 2008	85	121	546
# Matched	35	83	139
% Matched	50%	68.59%	29.26%

Now a matched dataset has been established, logistic regression is employed to determine what is associated with an individual staying with the programme. Once these estimates have been established, estimates of the number of individuals who remain in the programme are estimated, using this matched set as a base.

6. LONGITUDINAL ANALYSIS

6.1 Logistic Regression

When a model's dependent variable is Bernoulli $y \sim \text{Bernoulli}(p)$, the response is either 0, 1 and the mean μ is the probability p , the logistic regression model is an appropriate generalised linear model to use.

The logistic regression consists of a logit link function. The logit link is defined as:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \mathbf{x}^T \beta \quad (6.1)$$

The probability given by the model is defined as

$$p(x) = \frac{e^{(\mathbf{x}^T \beta)}}{1 + e^{(\mathbf{x}^T \beta)}} \quad (6.2)$$

where the probability always lies between 0 and 1.

A simple logistic regression model consists of a single continuous explanatory variable, and can be expressed as:

$$\mathbf{x}^T \beta = \alpha + \beta x = \log\left(\frac{p(x)}{1-p(x)}\right), \quad (6.3)$$

and with multiple explanatory variables, the model becomes:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_K x_K. \quad (6.4)$$

6.2 Model Selection

6.2.1 Criterion-based procedure

There are many criteria based procedures. Two of the most common are the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC).

The AIC is defined as:

$$-2 \max \log\text{-likelihood} + 2p, \quad (6.5)$$

where p is the number of parameters used in the model.

The smaller the AIC the better the model. The best model is one with a balance of the fit of model and the number of parameters used. The model with the smallest AIC is considered the best fit model.

The BIC is defined as:

$$-2 \max \log\text{-likelihood} + p \log n. \quad (6.6)$$

p is the number of parameters used in the model, and n is the sample size.

The BIC is very similar to AIC, the difference is large models (many parameters) are penalised more strongly by BIC. The key difference between AIC and BIC is the addition of $2p$ for AIC, and $p \log n$ for BIC.

Example

For $n = 7$ the AIC is equivalent to $2p$, while the BIC is equivalent to $1.95p$. When $n = 8$ the AIC is still equivalent to $2p$, while BIC is equivalent to $2.08p$. This means that for models with $n > 8$ BIC penalises the model more.

6.2.2 Testing-based procedure

There are 3 key testing based procedures:

- Backwards stepwise selection;
- Forward Stepwise selection;
- Mixed Stepwise selection.

Backwards stepwise selection starts with all possible explanatory variables, and then one by one the explanatory variables are removed based on a criteria selection basis. This criteria selection can be p -value, AIC, or BIC. This procedure is repeated until the removal of the variable brings about a worse criteria selection.

Forward stepwise selection is the converse of the backwards stepwise selection. It starts with no explanatory variables, and adds one by one until the addition of the variable brings a worse criteria selection.

The mixed stepwise selection is a combination of the backwards and forward stepwise selection. At each stage, explanatory variables can be added or removed to improve the selection criteria.

6.3 Logistic Regression Analysis

Using the analysis from the probabilistic matching the model is defined by:

$$y_i = \begin{cases} 1, & \text{if individual } i \text{ is matched} \\ 0, & \text{if individual } i \text{ is not matched} \end{cases} \quad (6.7)$$

The explanatory variables are, *NA Time*, *Clean Time*, *Sex*, *Age*, *Ethnicity*, *Urban*, *Education*, *Most Drug*, *Income Source*, *Drug Time*, *Meeting Frequency*, *Sponsor*, *Start Steps*, *All Steps*, and *Criminal*, for the base logistic regression model.

For all individuals who did not state their *NA Time* their *Clean Time* was imputed as their *NA Time*. Both *NA Time* and *Clean Time* have been grouped into the groups; 1-(5) years, 5-(10) years, 10-(15) years, 15-(20) years, and 20+ years.

Age has been grouped into < 30 years, 30-39 years, 40-49 years, and 50+ years. *Ethnicity* has been condensed into a binary variable of Maori, and Non-Maori.

The *Urban* variable is created by the original *City* variable where the major cities have been classified as *Urban*, while the minor cities and towns have been classified as rural. The major cities are Auckland, Christchurch, Dunedin, Hamilton, Napier, Rotorua, Tauranga, and Wellington.

Meeting Frequency variable is grouped into three different groups; Less than Once a week, weekly, and More than once weekly. *Most Drug* variable has been condensed into the three most common drugs; Alcohol, Cannabis, Opiates, and the rest are classified as Other. *Drug Time* variable is grouped into the periods of; Less than 1 year, 1-4 years, 5-9 years, 10-14 years, and 15 years or more.

Sponsor, *Start Steps*, *All Steps*, and *Criminal* and all binary variables which either take the value of yes or no.

Results

This logistic regression model has the baseline reference of a Non-Maori male less than 30 years old with NA Time and Clean Time less than 5 years. In an Urban city with No Education who is employed with Drug Time less than a year with the cannabis as preferred drug. This baseline person has a sponsor, a criminal record, and started and completed the 12 steps. This model has an AIC of 499.62.

Tab. 6.1: Summary of full logistic regression model

Variable	Level	Estimate	Std. Error	Pr(> Z)
Intercept		-12.9965	882.7440	0.9883
NA Time	5-(10)	-0.3935	0.4941	0.4258
	10-(15)	0.0353	0.4771	0.9410
	15-(20)	0.0858	0.6067	0.8875
	20+	-0.8576	0.9695	0.3764
Clean Time	5-(10)	0.9045	0.4593	0.0489
	10-(15)	0.4711	0.5399	0.3829
	15-(20)	0.6712	0.7231	0.3533
	20+	2.0875	1.2866	0.1047
Sex	Female	-0.0824	0.2602	0.7514
Age	30-39	0.1450	0.4045	0.7200
	40-50	-0.0333	0.4939	0.9462
	50+	0.2236	0.6078	0.7130
Ethnicity	Maori	-0.2753	0.3617	0.4466
Urban	Rural	-0.4598	0.3780	0.2239
Education	School	0.1509	0.4108	0.7133
	Vocational	-0.0488	0.4353	0.9107
	Tertiary	-0.2781	0.4103	0.4979
	Graduate	-0.1696	0.4276	0.6916
Most Drug	Alcohol	0.0958	0.3780	0.8000
	Opiates	0.4951	0.3764	0.1884
	Other	-0.2683	0.3379	0.4271
Income Source	Beneficiary	-0.6974	0.3142	0.0265
	Student	-1.9383	0.7907	0.0142
	Unpaid/Other	-1.2243	0.6232	0.0495
DrugTime	1-4 years	13.0678	882.7438	0.9882
	5-9 years	13.0186	882.7437	0.9882
	10-14 years	12.7272	882.7437	0.9885
	15 years or more	12.7533	882.7436	0.9885
Meeting	Less Than			
Frequency	Once A Week	-0.7600	0.5431	0.1617
	Weekly	-0.0776	0.3196	0.8082
Sponsor	No	0.1200	0.3219	0.7092
Start Steps	No	0.8160	0.4428	0.0653
All Steps	No	-0.4688	0.3162	0.1381
Criminal	No	-0.2302	0.2751	0.4026

Using the backwards stepwise method the model reduces down first by removing *Drug Time* providing an AIC of 492.61. Next *NA Time* is removed providing an AIC of 486.21, *Education* is removed after decreasing the AIC to 480.23, *Age* then *Clean Time* are removed with the reduction of AIC to 474.88 and 472.79 respectively. *Meeting Frequency* is removed after decreasing the AIC to 470.53. Next *Sponsor* is removed and AIC now becomes 468.67. *Criminal* and *Sex* are removed with AIC decreasing to 466.82, and 465.15. Next *Urban* is removed and AIC decreasing to 464.17, Finally *Ethnicity* is removed with a resulting AIC of 463.39. Removing any more explanatory variables will cause an increase in the AIC providing a worse model.

As a result the logistic regression now shows that being matched between the two surveys is explained by *Most Drug*, *Income Source*, *Start Steps*, and *All Steps* variables.

Tab. 6.2: Reduced logistic regression model summary

Variable	Levels	Estimate	Std. Error	Pr(> Z)
Intercept		-0.2502	0.2700	0.3540
Most Drug	Alcohol	0.1689	0.3492	0.6287
	Opiates	0.6680	0.3352	0.0463
	Other	-0.0775	0.3129	0.8045
Income Source	Beneficiary	-0.7699	0.2673	0.004
	Student	-1.9887	0.7635	0.0092
	Unpaid/Other	-1.1528	0.5875	0.0498
Start Steps	No	0.7692	0.4207	0.0675
All Steps	No	-0.6820	0.2559	0.0077

This model has a resulting AIC of 463.39.

Using this model individuals who have taken opiates the most are more likely to be matched in the next survey. Individuals that are not considered employed (Beneficiary, Student, Unpaid/Other) are significantly less likely of being matched in the next survey. Individuals that have completed all the 12 steps have a higher likelihood of being matched compared to those who have completed the 12 steps.

Using this logistic regression model the fitted probabilities are calculated for each individual in 2004. Using these fitted probabilities the dataset is broken into three unique groups low probability of being matched, moderate probability of being matched, and high probability of being matched. In the next chapter comparisons between the high probability and low probability

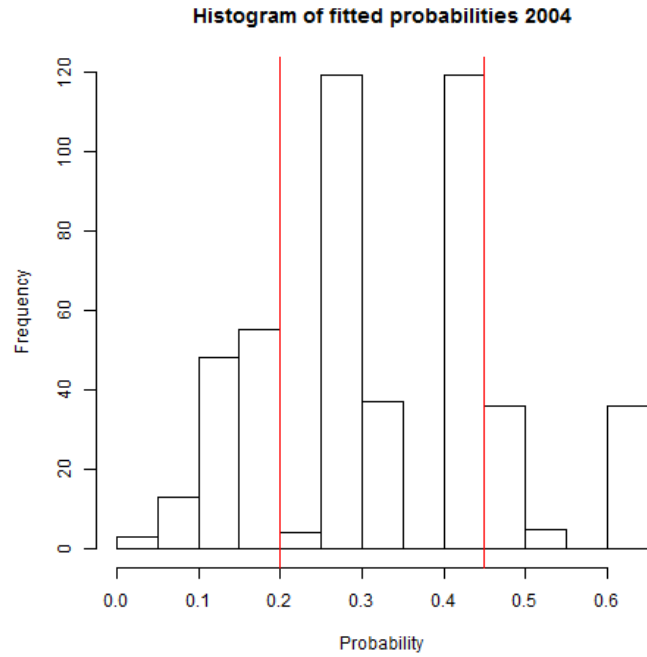


Fig. 6.1: Histogram of fitted probability groups

groups are made. These comparisons are used to determine characteristics about the two groups and determine what makes a person more likely to be matched in the program.

We see in Figure 6.1 the low probability group is defined as individuals with less than a 0.2 probability of being both years. Individuals with probability higher than 0.45 are defined as the high probability group.

This will be investigated later in chapter 7.

6.4 Longitudinal Analysis

Recall from Chapter 3 the simulation method of the population size estimates we now extend this methodology to the persisting population to estimate the population size of the persisting population.

At time t an individual can either be present (the individual is in the NA programme even if they do not attend meetings on a regular basis) or absent from the population. They can attend (the individual is participating in the

NA programme and is attending meetings) a meeting or not, they can be compliant (an individual is compliant if they are willing to participate in the survey if one is presented to them) or not, and they can respond or not.

If an individual is not present this implies they are not attending, and not responding. An individual is responding if and only if they are attending and are compliant.

$$I_{it} = \begin{cases} 1, & \text{if individual } i \text{ is present at time } t \\ 0, & \text{Otherwise} \end{cases} \quad (6.8)$$

The entire universe U is the (abstract) collection of all individuals at all time. The population at time t is defined as the set of all individuals present at time t

$$U_t = \{i : I_{it} = 1\}. \quad (6.9)$$

This population is broken down into two subgroups; those who attend meetings during survey week, and those who are absent from the meetings during the survey week.

$$a_{it} = \begin{cases} 1, & \text{if individual } i \text{ is attending meeting at time } t \\ 0, & \text{Otherwise} \end{cases} \quad (6.10)$$

At time t an individual may be complaint or not.

$$v_{it} = \begin{cases} 1, & \text{if individual } i \text{ is compliant at time } t \\ 0, & \text{Otherwise} \end{cases} \quad (6.11)$$

If an individual is compliant this determines if the individual responds, if present and attending. Responding individuals are defined as

$$r_{it} = \begin{cases} 1, & \text{if individual } i \text{ responds at time } t \\ 0, & \text{Otherwise} \end{cases} \quad (6.12)$$

An individual responding at time t must be present, attending, and compliant.

$$r_{it} = I_{it}a_{it}v_{it} \quad (6.13)$$

The size of the population at an given time t is the sum of all individuals present at time t i.e.

$$N_t = \sum_{i \in U} I_{it} \quad (6.14)$$

Recall from cross sectional estimates (Section 3.1.2) at any given time t , the probability that a responding individual attends is;

$$\pi_{it} = \min(1, g_{it}), \quad (6.15)$$

with the (Horvitz-Thompson) inverse probability weights of

$$w_{it} = \frac{1}{\pi_{it}} = \frac{1}{\min(1, g_{it})}. \quad (6.16)$$

The proportion of individuals whose attendance is their first attendance at time t is

$$\lambda_t = \frac{C_t}{C_t + \hat{H}_t}. \quad (6.17)$$

The proportion of the population that attends meetings in any given week at time t is

$$\Psi_t = \frac{C_t}{W_t} \quad (6.18)$$

The assumption of the attendance probability is the same for complying individuals and non-complying individuals. The proportion of the population that attends and is compliant is

$$\phi_t = \frac{\text{\#individuals attending and are compliant}}{\text{\#individuals attending}} = \frac{C_t + H_t}{A_t} \quad (6.19)$$

The odds for an individual being non-compliant at time t is:

$$\begin{aligned} \delta_t &= \frac{1 - \phi_t}{\phi_t} \\ &= \frac{1}{\phi_t} - 1 \\ \delta_t + 1 &= \frac{1}{\phi_t} \\ \frac{1}{\delta_t + 1} &= \phi_t. \end{aligned}$$

The estimated size of the population at time t is

$$\hat{N}_t = \frac{W_t A_t}{C_t + H_t} = \frac{W_t}{\phi_t}, \quad (6.20)$$

where W_t is the sum of the weights of the respondents at time t , which estimates the size of the compliant population at time t .

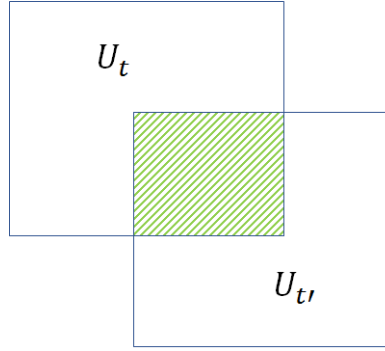


Fig. 6.2: Illustration of the persistent population

6.4.1 Persistent population estimate

The sets S_t and $S_{t'}$ are the sets of responding individuals from the first survey at time t and the second at time t' . The main goal is to determine how many individuals of the population at time t remain in at time t' .

U_t is the population at time t (size n_t) and $U_{t'}$ is the population at time t' (size $n_{t'}$). The shaded area in figure 6.2 is the persistent population (size $N_{tt'}$).

The probability that an individual i remains in the population from t to t' is

$$p_i = \Pr(I_{it'} = 1 | I_{it} = 1, \mathbf{x}_{it}), \quad (6.21)$$

where \mathbf{x}_{it} is a set of covariates at time t .

The probability that an individual i is matched depends on whether the individual responds at time t and time t' .

$$q_i = m\Pr(r_{it'} = 1 | r_{it} = 1, \mathbf{x}_{it}),$$

where m is the probability of a correct match with $0 < m \leq 1$. An individual responds if and only if they are present, they attend a meeting and they are compliant.

$$q_i = m\Pr(a_{it'} = 1, v_{it'} = 1, I_{it'} = 1 | a_{it} = 1, v_{it} = 1, I_{it} = 1, \mathbf{x}_{it})$$

This can be written as:

$$\begin{aligned}
 q_i &= m\Pr(a_{it'} = 1|v_{it'} = 1, I_{it'} = 1, a_{it} = 1, v_{it} = 1, I_{it} = 1, \mathbf{x}_{it}) \\
 &\quad \times \Pr(v_{it'} = 1|I_{it'} = 1, a_{it} = 1, v_{it} = 1, I_{it} = 1, \mathbf{x}_{it}) \\
 &\quad \times \Pr(I_{it'} = 1|a_{it} = 1, v_{it} = 1, I_{it} = 1, \mathbf{x}_{it}) \\
 &= m\Pr(a_{it'} = 1|I_{it'} = 1, \mathbf{x}_{it}) \\
 &\quad \times \Pr(v_{it'} = 1|v_{it} = 1) \\
 &\quad \times \Pr(I_{it'} = 1|I_{it} = 1, \mathbf{x}_{it})
 \end{aligned}$$

Four assumptions have been used to simplify the above expression:

1. Attendance depends only on presence/ absence and covariates;
2. current compliance depends only on previous compliance;
3. compliance status does not change $\Pr(v_{it'} = 1|v_{it} = 1) = 1$;
4. presence/absence is independent of attendance and compliance.

Recall that the probability that responding person i attends at time t is π_{it} (Equation 6.16). $\Pr(a_{it'} = 1|I_{it'}, \mathbf{x}_{it}) = \pi_{it'}$. Under the third assumption, $\Pr(v_{it'} = 1|v_{it} = 1) = 1$. By definition $p_i = \Pr(I_{it'} = 1|I_{it} = 1, \mathbf{x}_{it})$. The probability that an individual i is matched is defined as:

$$q_i = m\pi_{it'}p_i. \quad (6.22)$$

Thus the persistence probability, p_i , becomes:

$$p_i = \frac{q_i}{m\pi_{it'}} \quad (6.23)$$

$\pi_{it'}$ is set to the mean probability of attendance of respondents $\tau_{t'}$, at time t , where $\tau_{t'} = \frac{\sum_{i \in S_{t'}} \pi_{it'}}{C_{t'}}$. Therefore:

$$p_i \simeq \frac{q_i}{m\tau_{it'}}, \quad (6.24)$$

The average persisting probability is:

$$\bar{p} \simeq \frac{\bar{q}}{m\tau_{t'}} \quad (6.25)$$

where \bar{q} , the mean matched probability, and can be simply approximated by:

$$\bar{q} = \frac{M}{C_t}. \quad (6.26)$$

Where M is number of matched respondents at time t , and C_t is the number of respondents at time t .

A simple estimate of the number of individuals in the persisting population is

$$\begin{aligned}
 \hat{N}_{tt'} &= \hat{N}_t \bar{p} \\
 &= \frac{\hat{N}_t \bar{q}}{m\tau_{t'}} \\
 &= \frac{\hat{N}_t M}{C_t m\tau_{t'}} \\
 &= \frac{W_t M}{\phi_t C_t m\tau_{t'}} \\
 &= \frac{M}{\phi_t \Psi_t m\tau_{t'}}.
 \end{aligned} \tag{6.27}$$

A better approximation than equation 6.27, can be derived using the earlier work done with logistic regression. Using logistic regression q_i can be estimated for all individuals of the responding sample at time t .

Where \mathbf{x}_{it} is missing, the fitted value \hat{q}_i is estimated by \bar{q} , where \bar{q} is the mean of the fitted values observed.

The improved estimate of the persisting population becomes

$$\begin{aligned}
 \hat{N}_{tt'} &= \sum_{i \in S_t} \frac{\hat{p}_i}{\pi_{it} \phi_t} \\
 &= \sum_{i \in S_t} \frac{\hat{q}_i}{m\tau_{t'} \pi_{it} \phi_t}.
 \end{aligned} \tag{6.28}$$

6.4.2 Standard errors of the estimate of the persisting population

To create the standard errors of the persisting population $N_{tt'}$ we use a similar simulation method to that in chapter 3 where synthetic populations U_t^* and $U_{t'}^*$ are generated and observed. These synthetic populations are generated multiple times and estimates of \hat{N}_t^* , $\hat{N}_{t'}^*$, and $\hat{N}_{tt'}^*$ are recorded at each iteration.

To create the population estimates at time t a compliant set and a non-compliant set are generated. The compliant set is generated by generating $w_{it} = \frac{1}{\pi_{it}} = \frac{1}{\text{Prob. of attending}}$ copies of all individuals i in the respondent set S_i . All these generated individuals have $v_{it}^* = 1$.

When an individual has a non-integer w_{it} these individuals are generated by first generating the integer part of w_{it} . Then an additional generation would happen for individual i with a probability of $w_{it} - \text{int}(w_{it})$. For example an individual with $w_{it} = 3.46$ will have 3 copies generated and have a probability of 0.46 of having a fourth.

The non-compliant set is generated by taking $w_{it}\delta_{it}$ copies of individual i , where δ_t is the odds of being non-compliant at time t .

$$\delta_t = \frac{1 - \phi_t}{\phi_t} = \frac{1}{\phi_t} - 1 = \frac{A_t}{C_t + H_t} - 1.$$

All these generated individuals have $v_{it}^* = 0$. Combining these two sets provides the synthetic population U_t^* .

Using this synthetic population, the synthetic attendance for each individual a_{it}^* is also generated. The synthetic attendance is generated by using the probability of attendance, π_{it}^* , where $a_{it}^* \sim \text{Bernoulli}(\pi_{it}^*)$.

The synthetic response population S_t^* is the synthetic individuals who are compliant and attends the meeting $r_{it}^* = a_{it}^*v_{it}^*$. Using this information the synthetic population estimates at time t are

$$\begin{aligned} C_t^* &= \sum_{i \in U_t^*} r_{it}^* = \sum a_{it}^* v_{it}^* \\ A_t^* &= \sum_{i \in U_t^*} a_{it}^* \max(1, g_{it}^*) \\ H_t^* &= \sum_{i \in U_t^*} a_{it}^* \max(1, g_{it}^*) v_{it}^* - C_t^* \\ W_t^* &= \sum_{i \in U_t^*} w_{it}^* \\ \hat{N}_t^* &= \frac{W_t^* A_t^*}{C_t^* + H_t^*} \end{aligned}$$

In order to create the synthetic persistent population $N_{tt'}^*$ the synthetic populations at time t and t' must be generated and observed. For all individuals at time t the matched observations are also included in the synthetic population at time t .

When calculating $\hat{N}_{tt'}$ logistic regression is used on the responding population at time t to calculate the probability an individual is matched, \hat{q}_i . This calculation must be reflected in the synthetic population.

For each individual $i \in U_t^*$ there is the matched variable Y_i^* and the new synthetic matched set is generated by $Y_i^* r_{it}^*$ where the size of the matched set is $M^* = \sum_{i \in U_t^*} Y_i^* r_{it}^*$. Using this matched set the estimates of \hat{q}_i^* are estimated using the same logistic regression model defined using variables in Table 6.3.

The persistent synthetic population is estimated by:

$$\hat{N}_{tt'}^* = \sum_{i \in S_t^*} \frac{\hat{q}_i^*}{m^* \tau_{t'}^* \pi_{it}^* \phi_t^*}$$

where

$$\begin{aligned} m^* &= \frac{\sum_{i \in S_t^*} Y_{it}^*}{C_t^*} \\ \tau_{t'}^* &= \frac{\sum_{i \in S_{t'}^*} \pi_{it'}^*}{C_{t'}^*} \\ \pi_{it}^* &= \min(1, g_{it}^*) \\ \phi_t^* &= \frac{C_t^* + H_t^*}{A_t^*} \end{aligned}$$

This simulation and observation is repeated a large number of times e.g. 1000 times. The standard errors for the estimate $\hat{N}_{tt'}$ are the standard deviation for the set of estimates of the synthetic populations.

6.4.3 Longitudinal Estimates

To estimate the persistent population $\hat{N}_{tt'}$, first m_t , $\tau_{t'}$, π_{it} , ϕ_{it} and \hat{q}_i must be computed, where m is the probability that an individual at time t will be matched at time t' . We estimate this at 0.9, this comes from the quality of the matching variables from Chapter 5. The estimates A_t , H_t , and C_t are derived in Chapter 3, $A_t = 1172$, $H_t = 422$, and $C_t = 475$.

The estimated proportion of the population that is compliant ϕ_t is:

$$\phi_t = \frac{C_t + H_t}{A_t} = \frac{475 + 422}{1172} = 0.7653 = 0.77$$

The estimate for $\tau_{t'}$ is calculated using

$$\tau_{t'} = \frac{\sum_{i \in S_{t'}^*} \pi_{it'}}{C_{t'}} = \frac{532.92}{546} = 0.9761 = 0.98.$$

Using all these estimates, the persistent population is:

$$\begin{aligned}
 \hat{N}_{tt'} &= \sum_{i \in S_i} \frac{\hat{q}_i}{m\tau_{t'}\pi_{it}\phi_{it}} \\
 &= \sum_{i \in S_t} \frac{\hat{q}_i}{0.9 \times 0.98 \times \pi_{it} \times 0.77} \\
 &= \sum_{i \in S_t} \frac{\hat{q}_i}{0.67 \times \pi_{it}} \\
 &= 255.76 \\
 &\simeq 256
 \end{aligned}$$

To create standard errors, the datasets were simulated 1000 times and at each iteration these same estimates were created. Once all the 1000 iterations were completed, the standard deviation of the estimated $\hat{N}_{tt'}^*$ is the estimated standard error for $\hat{N}_{tt'}$.

In the first iteration the responding population C_t^* increased by an individual to 476. The attending population was $A_t^* = 1521.52$, $H_t = 720.8$. The proportion of the population that is compliant is $\phi_t^* = 0.7866$. The probability of a correct match in the first iteration $m = 0.2983$. The mean probability of attendance amongst the respondents at time t' is $\tau_{t'} = 0.9729$. The estimate of the persistent of population in the first iteration, $\hat{N}_{tt'}^* = 794.868$.

The standard deviation of the simulated persistent population is $SD(\hat{N}_{tt'}^*) = 28.06743$. The estimates of the persistent population is summarised in Table 6.4.3.

Tab. 6.3: Summary information of persistent population

	Estimate	Std. Error
ϕ_t	0.77	0.01
$\tau_{t'}$	0.98	0.01
Z_t	274	28.97
$N_{tt'}$	256	8.34
N_t	695	(533, 840)
$N_{t'}$	771	(592, 836)

6.4.4 *Concluding Remarks*

Using the theory established throughout this chapter the size of the persistent population is 256. After running a 1000 iteration simulation study the standard error on this estimate is 8.34. The persistent population size estimate has a 95% confidence interval of (239.65, 272.34).

Over the four year period, there is a retention rate of 37%, giving an annual retention rate of 78%.

7. ANALYSIS OF MATCHED DATA

7.1 *Probabilistic Matching*

Recall from Chapter 5 we started with a matching based on individuals with a long clean time. This was done with the assumption the individuals with long clean time showed more stable behaviour and were less likely to relapse and hence more likely to be matched. Using information gained from this group, we then translated the same analysis application to a group with slightly longer clean time. Using this group we strengthened our assumptions made in the analysis. Using these strengthened assumptions, we then applied the matching analysis to all the individuals available in the dataset. In the end we were able to match 30% of the 2004 population.

We have seen individuals who are employed, and have higher education tend to have longer clean times. Using the matched dataset we can take this analysis a step further and investigate changes between matched individuals.

The areas that we are interested in are, if individuals can remain employed, increase their education status, if any of the individuals develop health problems, and if any individuals have a criminal record.

The majority of individuals who were employed at the time of the first survey after 4 years in the programme kept their employment status, with a small portion however losing their employment status. Two thirds of the unemployed individuals did manage to gain an employment status after just 4 years of the programme.

Tab. 7.1: Matched comparison of employment status

		2008	
		Employed	Unemployed
2004	Employed	88.1%	11.9%
	Unemployed	57.9 %	42.1%

The majority of individuals remained at the same education level, a small

proportion do change their level as seen in Table 7.2. From the small proportion that actually do change their level after remaining in the programme most shift up into the next level of education; e.g. none to school. Due to the surveys being four years apart individuals are capable of going from having no education to having a tertiary level of education.

Some individuals also shift down levels. This is technically impossible, and may be due to respondents not remembering correctly, misinterpreting, or possibly due to being incorrectly matched because individuals could misinterpret the question.

Tab. 7.2: Matched comparison of education status

		2008 None	School	Vocational	Tertiary	Graduate	Did Not Respond
2004	None	71.4%	9.5%	9.5%	4.8%	0.0%	4.8%
	School	6.9%	82.8%	3.4%	3.4%	3.4%	0.0%
	Vocational	0.0%	4.2%	75.0%	1.6%	4.2%	0.0%
	Tertiary	3.2%	3.2%	3.2%	74.2%	16.2%	0.0%
	Graduate	0.0%	0.0%	2.9%	8.9%	88.2%	0.0%

Many of the matched individuals didn't answer the medical and mental question. Obviously people consider the question about their medical and mental state a bit invasive. The majority still replied to the question.

Individuals in the matched set display the trend of having no medical/mental conditions. This does not mean if an individual does have medical or mental conditions they will not be a match. As seen in table 7.3 and table 7.4 there is a small proportion of individuals who develop some condition even after remaining in the programme for four years.

Tab. 7.3: Matched comparison of Post Medical status

		2008 No	Yes	Did Not Respond
2004	No	72.0%	10.2%	17.8%
	Yes	31.3 %	37.5%	31.3%
	Did Not Respond	80.0%	0.0%	20.0%

We see that individuals can add a criminal record, but they cannot remove one. Certain individuals stated they had a criminal record in 2004 but not

Tab. 7.4: Matched comparison of Post Mental status

		2008 No	Yes	Did Not Respond
2004	No	81.9%	13.8%	4.3%
	Yes	46.7%	33.3%	20.0%
	Did Not Respond	75.0 %	12.5%	12.5%

one in 2008. Individuals such as these would likely have criminal records due to other crimes unrelated to drug and misinterpreted the question in the first survey.

Another proportion have ended up with a criminal record since the study (see Table 7.5). Because this group has already completed the survey, the assumption was made this group wouldn't mistake their class of conviction as they have already completed the survey. These individuals have ended up with a criminal record based on their drug use.

Tab. 7.5: Matched comparison of criminal status

		2008 No	Yes	Did Not Respond
2004	No	74.6%	23.8%	1.6%
	Yes	15.8%	84.2%	0.0%
	Did Not Respond	0.0%	0.0%	0.0%

7.2 Longitudinal Estimates

We applied a logistic regression model to the data. After the model was fitted, the fitted probabilities of being in the later survey were calculated for each individual. Once these probabilities were established the dataset was broken down into three groups, high probability to being matched, moderate probability to being matched, and low probability to being matched. The aim of this was to establish what parameters affect information about an individual staying/leaving the programme.

Note that, $p_i \neq q_i$, but rather $p_i \propto q_i$. So having a high or low q_i suggest that p_i is high/low.

After splitting the dataset into the three groups, the high probability, the moderate probability, and low probability were compared. The variables that showed differences between the two surveys were:

- Clean Time;
- Age;
- Education;
- Most Drug;
- Income Source;
- All Steps;

We have shown already that clean time is a big factor for an individual staying in the programme. It also has been shown that the older the individual the longer the individual remains in the programme. The parameters education and income source showed that higher an individuals education is and being employed individuals are more likely to be matched.

For the low probability group $\simeq 85\%$ of individuals in this group had clean time of less than 5 years, while in the high probability group $\simeq 69\%$ had a clean time of less than 5 years, and $\simeq 26\%$ had clean time of more than 5 years. Individuals with longer clean time are more likely to be matched than those with shorter clean time this is illustrated in Figure 7.1 where more individuals with a high probability of being persistent have a clean time of 5-10 years.

Individuals who were in the high probability group tended to be in the older age groups, while individuals in the low probability group fell into the younger age groups, with $\simeq 75\%$ individuals being under 40 years old. From Figure

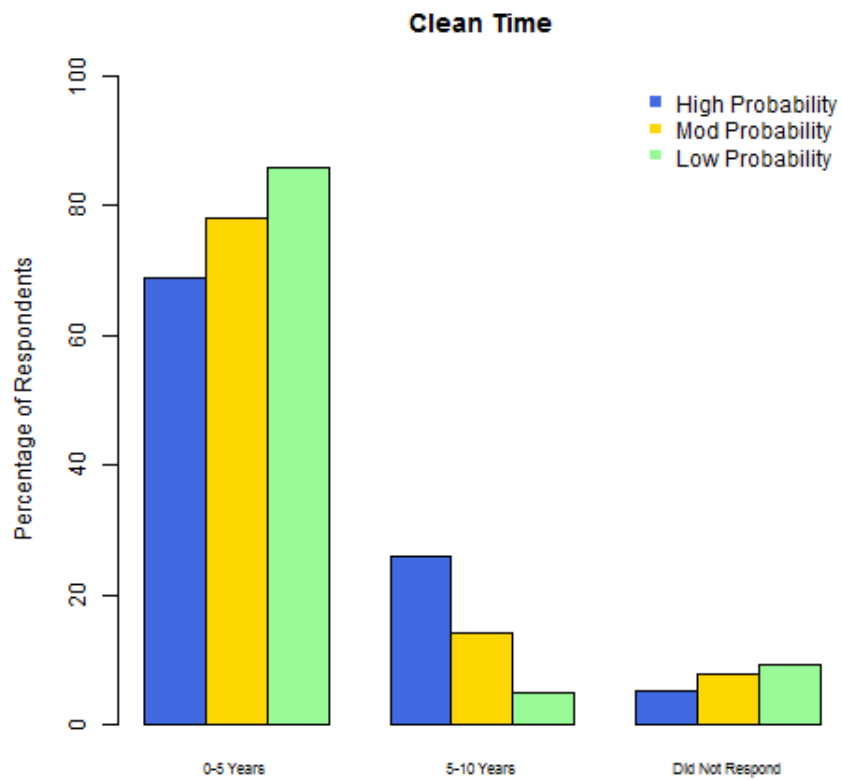


Fig. 7.1: Distribution of individuals clean time broken down into persisting population groups

7.2 the distribution of the high probability group tends towards being older, with the majority 40-49 years old. The moderate probability group, and the low probability group are distributed with lower age compared with the high probability group, with the majority of individuals with low and moderate probability of being persistent falling into the age group of 30-39 years.

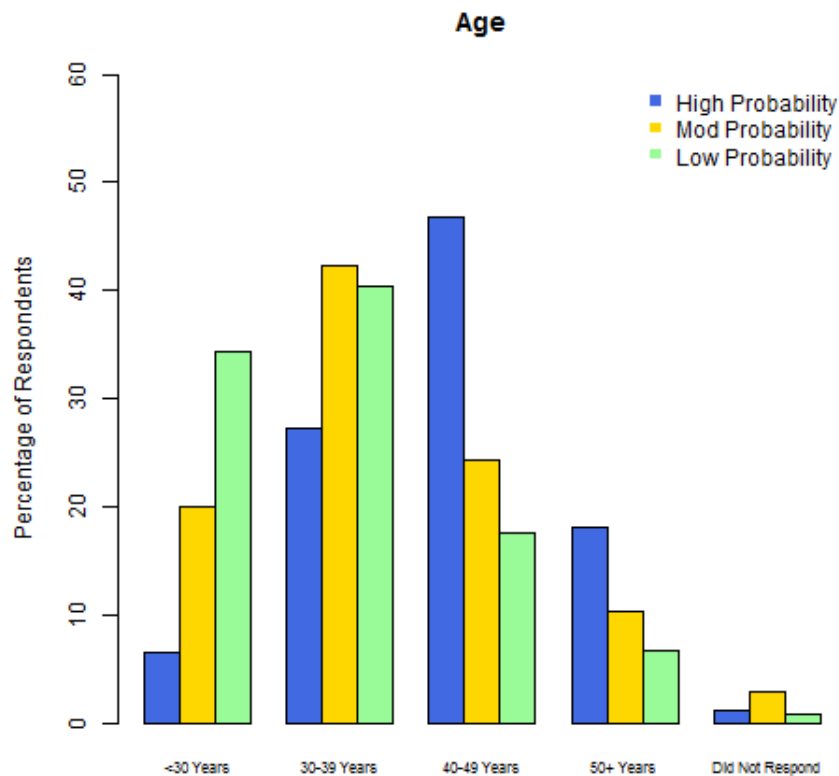


Fig. 7.2: Distribution of individuals age broken down into persisting population groups

An individuals level of education had little difference between the three groups. The difference between the two groups (high/low) was that individuals that were graduates were more likely to be in the high probability group (37.3%), and $\simeq 50\%$ of individuals with a low probability of being persistent have a vocational level of education, as illustrated in Figure 7.3.

Individuals who were in the high probability group had alcohol and opiates as their most used drug while individuals in the low probability group

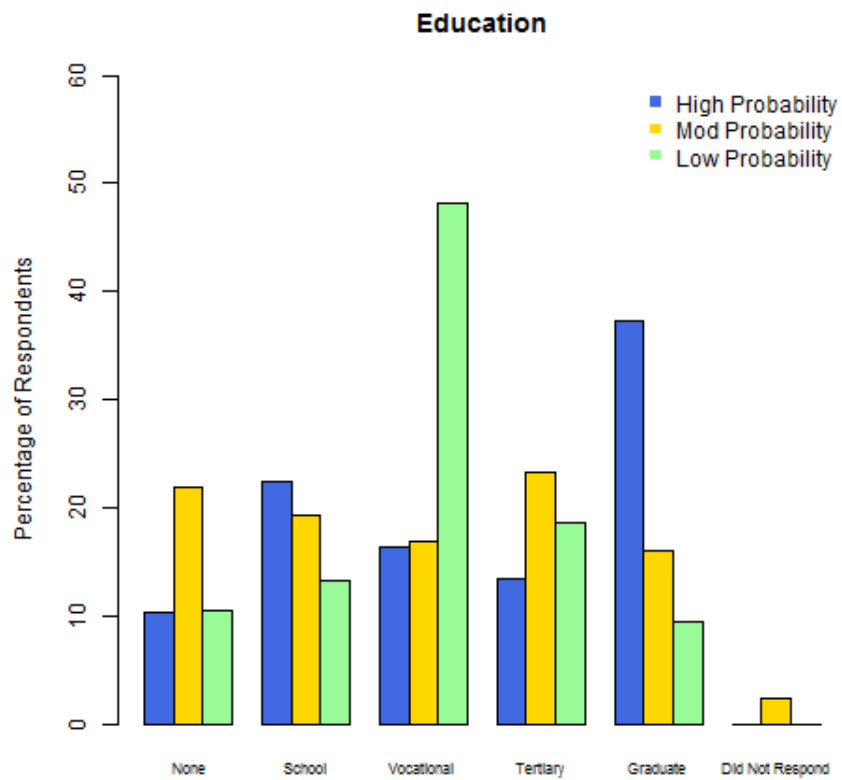


Fig. 7.3: Distribution of individuals education level broken down into persisting population groups

had “other” as their most used drug (42.9%). Individuals who took opiates (46.8%), or alcohol (46.8%) the most are more likely to be persistent, illustrated in Figure 7.4

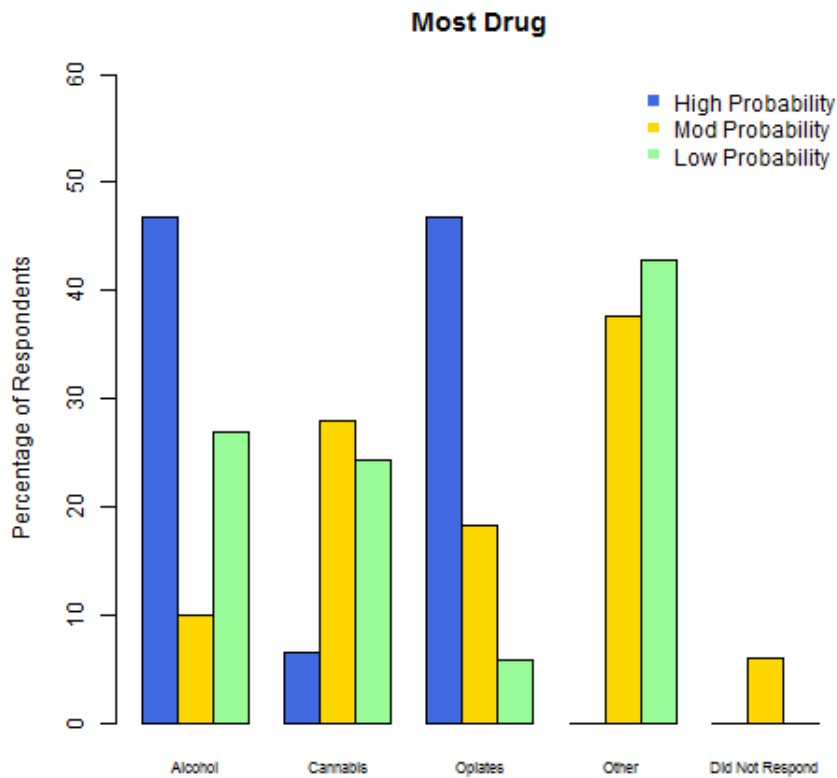


Fig. 7.4: Distribution of most drug an individual used broken down into persisting population groups

An individual’s source of income provides the strongest defining characteristic. An individual is likely to be persistent if they are employed, evidence of which is illustrated in Figure 7.5, where only individuals with a high probability are employed, and the majority of individuals with moderate probability of being persistent are employed, while no individuals with low probability of being persistent are employed. This shows that an individual’s source of income is an important factor to being persistent in the Narcotics Anonymous programme.

Individuals who have completed all the 12 steps is another area which provides a clear defining indication of being persistent. The majority of indi-

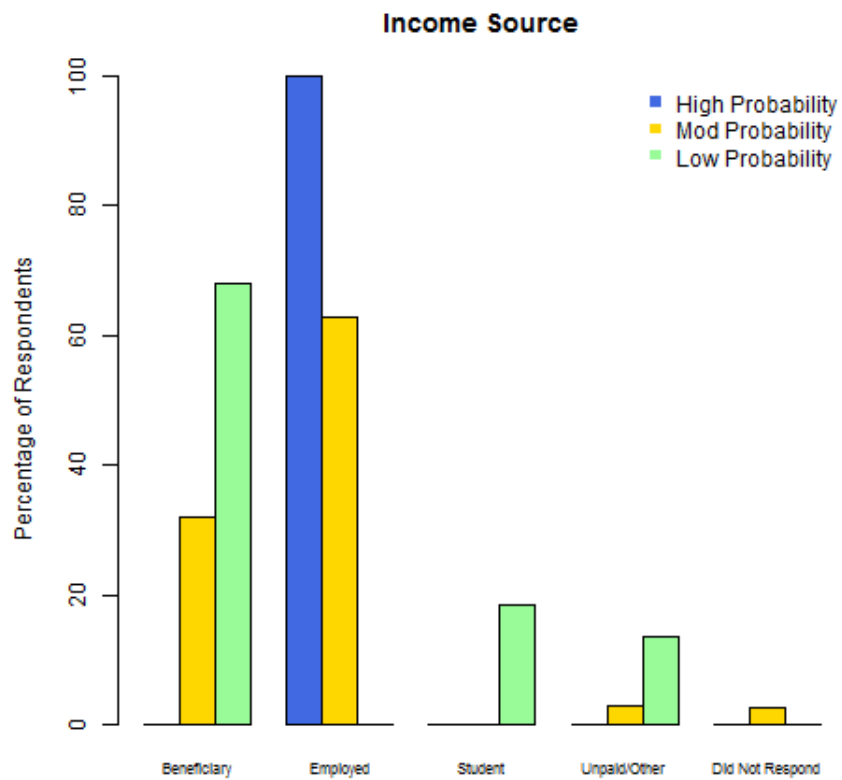


Fig. 7.5: Distribution of individual income source broken down into persisting population groups

viduals who are in the high probability group had completed all the 12 steps while individuals who are in the low probability group had not completed all the 12 steps, leading to the conclusion that individuals who have completed all the 12 steps are more likely to be persistent (Figure 7.6).

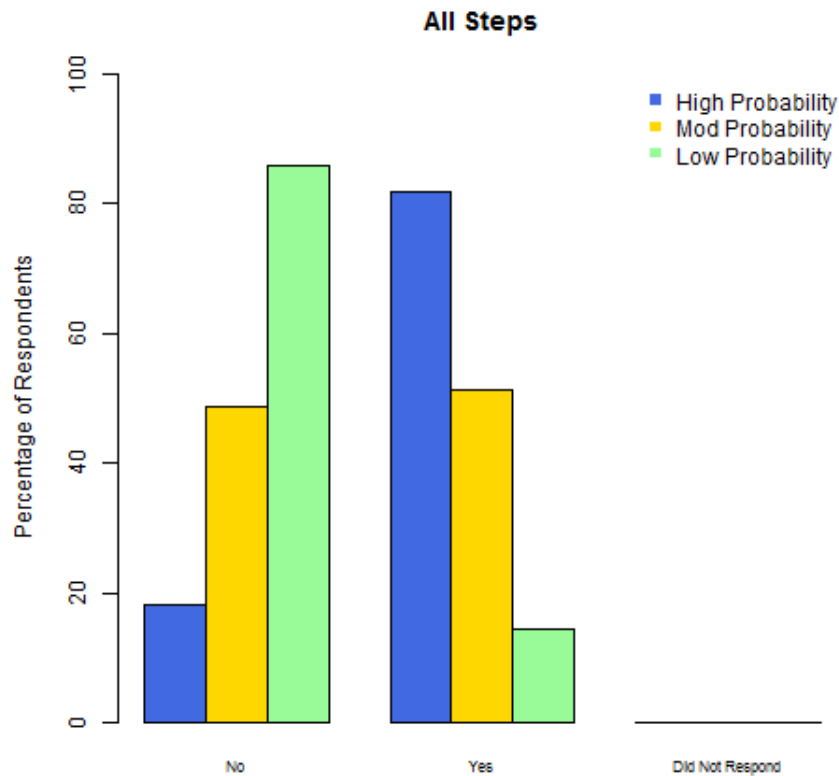


Fig. 7.6: Distribution of whether individual has completed the 12 steps of the programme broken down into persisting population groups

We have not covered an individual's level of education in the reduced logistic regression model, as the individual's income source looks related to the individual's level of income, as in Figure 7.5 we have seen that all individuals with high probability of persisting, and $\simeq 62\%$ of individuals with mid probability of persisting are employed. Individuals who had a high probability also favoured Alcohol and Opiates, and this would be tied into the individuals employment status, with those employed being able to avoid more costly drugs more often.

8. SUMMARY

8.1 *Results*

From cross sectional estimation (Chapter 3) the estimated number of individuals attending in any given week for 2004 is 620 (475, 742). We observed there were 475 individuals that answered during the survey week. We therefore estimate the additional 145 individuals in our estimate for attending the meeting as individuals in attendance but refusing to participate in the survey.

The total population for 2004 estimation is 695 (533, 840), individuals. This estimate is the estimate of the whole population rather those attending in any given week.

The number of individuals attending in any given week for 2008 is 650 (546, 739). The total population estimate for 2008 is 771 (592, 836).

These estimates provide insight into the population size of the Narcotics Anonymous program, but have provided little insight on the properties of the individuals who stay with or leave the program.

The variables that were of interest in this study were the age of the individual at the time of the survey, the individuals clean time and time spent with Narcotics Anonymous. Education and income source were also investigated as key points of interest.

For 2004 the average age of the individuals was 37 (36, 38) years old. In 2008 the average age increases to 39 (38,40) years old. There was a clearly defined increase in the average age of individuals between the two survey years.

The average clean time for 2004 was 4.83 (4.31,5.35) years, in 2008 this increased to 5.41 years.

A linear regression investigation of the 2004 and 2008 data was conducted. Estimates were created to establish the impact of age, sex, ethnicity, education, and income source on, clean time, and time in Narcotics Anonymous.

Bootstrap estimation was done to create the confidence intervals for the estimates.

From this linear regression we discovered that both time spent in NA and clean time are expected to be longer as the individual ages.

This is essentially showing the older an individual is the longer the expected clean time/ NA time is, compared to a younger individual.

Individuals who were employed at the time of the survey had longer clean time, and were part of the NA program, longer than individuals who were unemployed at the time of the survey.

Those individuals who are employed tend to have more structure in their lives than individuals who are unemployed.

Individuals who obtained their tertiary qualification (Graduate or Postgraduate qualification) demonstrated they could remain clean, or stick with the NA program. While individuals who did not gain any qualifications showed lower clean time and higher chance of relapsing.

Narcotics Anonymous has developed a program with an annual retention rate of 78%. This persistent rate was a real finding for this study. It suggests, individuals who are employed and have completed all the 12 steps are more likely to stay in the program.

8.2 *Further Research*

This study indicates with further research additional worthwhile information could be established, and key estimates improved.

For probabilistic matching this study only treats the variables the matching is based on as strictly independent of each other. Dependence between certain variables would require further research to ensure the weighting of matched pairs would not be biased, although some dependence would be expected. An example of expected dependence occurs between city and region, as if someone says they live in Auckland it would mean they are also in the Northern region.

The further research area around the dependence of variables would be based on variables with unknown correlations. Because all the weights have been calculated based on the sum of the variables used for matching, this means

if two variables are correlated then, these variables weights are biased due to this correlation.

This has a negative impact on the matching weights leading to bias in the matched pairs, either under estimation or over estimation in the number of matched pairs.

Another area of the probabilistic matching research that requires further study is the optimization of the algorithm that determines the variables to be used as a base for matching. In the research conducted a basic type algorithm approach was used where the best improvement was taken.

In probabilistic matching, individuals are matched based on weights which is the sum of all individual variable weights. The total possible combination of ways to calculate the overall weights is $2^k - 1$. This is infeasible when many variables are used. In this study, 35 variables were available to base the matching on resulting in 34,359,738,367 ways to arrange the variables for the calculation of the matching weights. Instead of searching through all possible combinations we used the algorithm to work through a feasible number of steps, however this algorithm always took the “most improved” path. Further research would be investigating methods for when a “poor” path is declared early on in the algorithm but is considered the “best” path at the end.

Along with determining the most efficient algorithm, the blocking technique established by Fellegi & Sunter (1969) would also be applied to this dataset in future studies. This blocking technique groups similar variables together reducing the complexity of the matching the dataset down.

Throughout this study the calculation of the weights considered all the matches as being true matches. Further research would consider some of the questionable declared matches as false matches and the implications that follow. False matches could explain some of the abnormal movement of some of the persistent individuals. e.g. having vocational education in 2004 and having no education in 2008.

8.3 Final Remarks

Drug use in New Zealand is always an ongoing issue with certain individuals becoming addicted to drugs, Narcotics Anonymous has established a programme, with the average clean time of an individual being proportional to their age.

Being persistent in Narcotics Anonymous was not determined by Sex or Ethnicity, but an individual's education, and employment status was a significant factor for an individual's clean time.

There were multiple significant factors for an individual to be matched between the two survey years. Some of these included sex, age, education and education prior to NA.

When investigating the probability of being in the persistent population individuals with high probability were older employed individuals with longer clean time, who had completed all the 12 steps.

Multiple statistical techniques have been used throughout this study. These included bootstrap estimation and linear regression estimation to establish estimates of cross-sectional population size and population dynamics. Probabilistic matching and logistic regression were used to establish estimates of persistent population between the two survey years.

This study builds on the base presented by Richard Arnold and Sharleen Forbes. A review of the cross sectional estimation was re-evaluated and refined using the bootstrap method. In addition to the cross sectional estimation, probabilistic matching was used to establish a matched set. This study has shown the base of probabilistic matching has provided strong evidence for future investigation. This study only dealt with the assumption that all variables are independent for the probabilistic matching. Future studies can build on this basis.

As the longitudinal estimates of individuals remaining, leaving, and joining NA between 2004 and 2008 have been established by using this probabilistic matching, any improvements made on the probabilistic matching estimates would also assist in improving the population estimates. These improvements provide clearer insights to population sizes and dynamics for any future investigations by the Narcotics Anonymous Fellowship.

BIBLIOGRAPHY

- Agresti, A. (2002), ‘Wiley series in probability and statistics’, *Analysis of Ordinal Categorical Data, Second Edition* pp. 397–405.
- Agresti, A. (2007), ‘Logistic regression’, *An Introduction to Categorical Data Analysis, Second Edition* pp. 99–136.
- Anonymous, N. (2005), ‘Aotearoa nz regional survey of narcotics anonymous members: “making your recovery count”’.
- Archer, K. J., Lemeshow, S. & Hosmer, D. W. (2007), ‘Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design’, *Computational Statistics & Data Analysis* **51**(9), 4450–4464.
- Ardilly, P. & Le Blanc, D. (2001), ‘Sampling and weighting a survey of homeless persons: a french example’, *Survey methodology* **27**(1), 109–118.
- Armstrong, J. & Mayda, J. (1992), ‘Estimation of record linkage models using dependent data’, *Proceedings of the Section on Survey Research Methodology. American Statistical Association* pp. 853–858.
- Arnold, R. & Forbes, S. (2005), ‘A method for estimating active membership for organisations that do not maintain registers of members’.
- Ayhan, H. & Ekni, S. (2003), ‘Coverage error in population censuses: the case of turkey’, *Survey Methodology* **29**(2), 155–166.
- Blakely, T. & Salmond, C. (2002), ‘Probabilistic record linkage and a method to calculate the positive predictive value’, *International journal of epidemiology* **31**(6), 1246–1252.
- Christensen, R. (2006), *Log-linear models and logistic regression*, Springer Science & Business Media.

- Daggy, J. K., Xu, H., Hui, S. L., Gamache, R. E. & Grannis, S. J. (2013), 'A practical approach for incorporating dependence among fields in probabilistic record linkage', *BMC medical informatics and decision making* **13**(1), 97.
- Dobson, A. J. & Barnett, A. (2002), 'An introduction to generalized linear models. 2002', *CRC Pr I Llc* .
- Dunn, H. L. (1946), 'Record linkage', *American Journal of Public Health and the Nations Health* **36**(12), 1412–1416.
- Efron, B. & Tibshirani, R. J. (1994), *An introduction to the bootstrap*, CRC press.
- Fellegi, I. P. & Sunter, A. B. (1969), 'A theory for record linkage', *Journal of the American Statistical Association* **64**(328), 1183–1210.
- Gelman, A. & Hill, J. (2006), *Data analysis using regression and multi-level/hierarchical models*, Cambridge university press.
- Gu, L., Baxter, R., Vickers, D. & Rainsford, C. (2003), 'Record linkage: Current practice and future directions', *CSIRO Mathematical and Information Sciences Technical Report* **3**, 83.
- Hill, S., Atkinson, J. & Blakely, T. (2002), 'Anonymous record linkage of census and mortality records: 1981, 1986, 1991, 1996 census cohorts'.
- Jaro, M. A. (1995), 'Probabilistic linkage of large public health data files', *Statistics in medicine* **14**(5-7), 491–498.
- Jeansonne, A. (2002), 'Loglinear models', *Retrieved June* **9**, 2009.
- John Lu, Z. (2010), 'The elements of statistical learning: data mining, inference, and prediction', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **173**(3), 693–694.
- Krewski, D., Dewanji, A., Wang, Y., Bartlett, S., Zielinski, J. & Mallick, R. (2005), 'The effect of record linkage errors on risk estimates in cohort mortality studies', *Survey Methodology* **31**(1), 13–21.
- Lawson, J., White, D., Price, B. & Yamagata, R. (2002), 'probabilistic record linkage for genealogical research', *Brigham Young University Studies* **41**(2), 161–174.

- Lohr, S. (2007), ‘Recent developments in multiple frame surveys’, *cell* **46**(42.2), 6.
- Lohr, S. (2009), *Sampling: design and analysis*, Nelson Education.
- Lumley, T. (2011), *Complex surveys: a guide to analysis using R*, Vol. 565, John Wiley & Sons.
- Machado, C. J. & Hill, K. (2004), ‘Probabilistic record linkage and an automated procedure to minimize the undecided-matched pair problem’, *Cadernos de Saúde Pública* **20**(4), 915–925.
- Mason, K. (2005), ‘Risking more than just money: Problem gambling in new zealand’.
- Mason, K. & Arnold, R. (2007), ‘Problem gambling risk factors and associated behaviours and health status: results from the 2002/03 new zealand health survey’, *The New Zealand Medical Journal (Online)* **120**(1257).
- Murray, J. S. (2016), ‘Probabilistic record linkage and deduplication after indexing, blocking, and filtering’, *arXiv preprint arXiv:1603.07816*.
- Newcombe, H. B., Kennedy, J. M., Axford, S. & James, A. P. (1959), ‘Automatic linkage of vital records’, *Science* **130**(3381), 954–959.
- Pledger, S., Pollock, K. H. & Norris, J. L. (2003), ‘Open capture-recapture models with heterogeneity: I. cormack-jolly-seber model’, *Biometrics* **59**(4), 786–794.
- Pledger, S., Pollock, K. H. & Norris, J. L. (2010), ‘Open capture-recapture models with heterogeneity: II. jolly-seber model’, *Biometrics* **66**(3), 883–890.
- Prayer, S. (1976), ‘Na white booklet, narcotics anonymous’.
- Presnell, B. (2000), ‘An introduction to categorical data analysis using r’.
- Sariyar, M. & Borg, A. (2010), ‘The recordlinkage package: Detecting errors in data’, *The R Journal* **2**(2), 61–67.
- Singh, K. & Xie, M. (2008), ‘Bootstrap: a statistical method’, *Unpublished manuscript, Rutgers University, USA. Retrieved from <http://www.stat.rutgers.edu/home/mxie/RCPapers/bootstrap.pdf>*.

-
- Tromp, M., Meray, N., Ravelli, A. C., Reitsma, J. B. & Bonsel, G. J. (2008), 'Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies', *Journal of the American Medical Informatics Association* **15**(5), 654–660.
- Winkler, W. E. (1992), Comparative analysis of record linkage decision rules, in 'Proceedings of the section on survey research methods', pp. 829–834.
- Winkler, W. E. (1995), 'Matching and record linkage', *Business survey methods* **1**, 355–384.
- Winkler, W. E. (2000), 'Frequency-based matching in fellegi-sunter model of record linkage', *Bureau of the Census Statistical Research Division* **14**.
- Yancey, W. E. (2000), Frequency-dependent probability measures for record linkage, in 'Proceedings of'.
- Zhu, R., Zhang, J., Zhang, D. & Yan, G. (2010), 'Stepwise variable selection for loglinear mixture in record linkage', *European Journal of Pure and Applied Mathematics* **3**(2), 141–162.

APPENDIX

A. QUESTION AND SURVEY FORMS

This appendix contains the survey forms for 2004 and 2008 participants, along with the instructions sent on how to run the survey.

NA Member Survey: Appendix 1

AOTEAROA NEW ZEALAND MEMBER SURVEY



About You

1. Please indicate how much clean time you currently have?

1. What is your clean time date __ DD/MM/YY __
2. When did you first attend NA __ MM / YY __

2. Gender

1. ☐ Male
2. ☐ Female

3. My age is _____ years

4. What is the closest city or town to where you live?

5. What is the main ethnic group you belong to? (Tick *one* only)

1. ☐ New Zealand European / Pakeha
2. ☐ Maori
3. ☐ Samoan
4. ☐ Cook Island Maori
5. ☐ Tongan
6. ☐ Niuean
7. ☐ Fijian
8. ☐ Chinese
9. ☐ Indian
10. ☐ Other: _____

About your recovery

6. Do you have a sponsor now?

1. ☐ Yes, I speak with or contact my sponsor:
☐ daily ☐ weekly ☐ other _____
2. ☐ No I don't have a sponsor.

7. Do you sponsor others now?

1. ☐ Yes, I sponsor _____ people.
2. ☐ No

8. How often do you go to NA meetings now? (Complete *one* line only)

1. _____ times a week
2. _____ times a month

9. Where did you get clean? (tick *one* only)

1. ☐ NA 2. ☐ Treatment. 3. ☐ Other
4. ☐ Other 12 Step fellowship please specify _____

10. Have you started to work the 12 steps?

1. ☐ Yes
2. ☐ No

11. Have you worked on all the 12 steps at least once?

1. ☐ Yes
2. ☐ No

12. What NA service have you been involved with? (Tick *all* that apply)

1. ☐ None
2. ☐ Group service: (secretary, treasurer, coffemaker, literature person, etc.)
3. ☐ Area service (inc. sub-committees & GSR)
4. ☐ Regional service
5. ☐ World service
6. ☐ Other _____

13. Do you currently attend meetings of any other 12 step fellowships?

1. ☐ Yes
2. ☐ No

14. Do you currently attend other types of addiction counselling or addiction support programs?

1. ☐ Yes
2. ☐ No

15. Please indicate which of the following influenced you the most to come to your first NA meeting. (Tick *one* only)

1. ☐ NA member
2. ☐ NA literature / phoneline / flyer / website
3. ☐ Treatment facility / health care agency
4. ☐ NA meeting in prison / corrections facility
5. ☐ AA member or group
6. ☐ Family / friend / Nar-Anon member
7. ☐ Self referred
8. ☐ Court / probation or parole officer
9. ☐ Employer or co-worker
10. ☐ Newspaper, magazine, radio, or TV
11. ☐ Counsellor, teacher, clergy member
12. ☐ Other: _____

Before you came into recovery

16. How long did you use drugs for?

1. ☐ under 1 year
2. ☐ Between 1 year and 4 years
3. ☐ Between 5 years and 9 years
4. ☐ Between 10 years and 14 years
5. ☐ 15 years and over

PLEASE TURN OVER >>

Fig. A.1: Page 1/2 2004 survey questionnaire

17. What drugs did you use on a regular basis:
(Tick *all* that apply)

1. ☐ Alcohol
2. ☐ Cannabis (pot, hashish, etc.)
3. ☐ Methamphetamine (P etc.)
4. ☐ Cocaine
5. ☐ Barbiturates (downers, etc.)
6. ☐ Tranquillisers (Valium, etc.)
7. ☐ Hallucinogens (LSD, etc.)
8. ☐ Inhalants (glue, etc.)
9. ☐ Opiates (heroin, etc.)
10. ☐ Other stimulants (speed, etc.)
11. ☐ Methadone

Other: _____

18. Did you have one drug of choice?

1. ☐ Yes, select number from the list above: _____
2. ☐ No

19. Was there one drug you used the most?

1. ☐ Yes, select number from the list above: _____
2. ☐ No

More about you**20. Main source of income now.** (Tick *one* only)

1. ☐ Full-time employment (30 hours or more)
2. ☐ Part-time employment (30 hours or less)
3. ☐ Self employment, full or part-time
4. ☐ Retired (Super Annuitant)
5. ☐ Student loan / grant
6. ☐ Private income
7. ☐ Unpaid worker (carer, homemaker etc)
8. ☐ Beneficiary (ACC, sickness, unemployed, DPB)

Other: _____

21. Main source of income before attending NA

Please select number from the list above: _____

22. What is your highest educational qualification now?

1. ☐ No qualification
2. ☐ School qualification
3. ☐ Trade / vocational qualification
4. ☐ Tertiary qualification
5. ☐ Graduate or postgraduate qualification

23. What was your highest education level prior to attending NA?

Please select from the list above: _____

24. Main type of paid work you're doing now.
(Tick *one* only)

1. ☐ Unskilled work / labouring
2. ☐ Clerical work
3. ☐ Service / sales work / hospitality
4. ☐ Trades worker
5. ☐ Agriculture / fishery
6. ☐ Plant / machine operator
7. ☐ Craft worker / artist / musician / actor
8. ☐ Technical
9. ☐ Manager/administrator/legislator/professional
10. ☐ Health professional
11. ☐ Student
12. ☐ None

Other: _____

25. Main type of paid work prior attending NA

Please select number from the list above: _____

Other Information**26. Before your clean date, did you contract / develop a diagnosed health condition?**

1. ☐ No
2. ☐ Medical (eg. Hep. C, HIV, Organ Damage)

Please Specify _____

3. ☐ Mental Illness (Bipolar/Depression/Psychosis)

Please specify _____

27. Do you still experience the symptoms that resulted in your diagnosis?

1. Yes, ☐ Medical ☐ Mental
2. ☐ No

28. Since your clean date, have you contracted / developed a diagnosed health condition?

1. ☐ No
2. ☐ Medical (eg. Hep. C, HIV, Organ Damage)

Please Specify _____

3. ☐ Mental Illness (Bipolar/Depression/Psychosis)

Please specify _____


29. Do you have criminal convictions as a result of your drug use?

1. ☐ Yes
2. ☐ No

Thank You for your Service

NOT COLLECTED? If for any reason this form has not been collected at the meeting please post it to Survey Work Group, PO Box 9051, Wellington by the 29th of November at the latest.

Fig. A.2: Page 2/2 2004 survey questionnaire

AOTEAROA NEW ZEALAND MEMBER SURVEY 	
About You	
1. Please indicate how much clean time you currently have? (DD or MM not required if unsure) 1. What is your clean time date ___ DD / MM / YY ___ 2. When did you first attend NA ___ DD / MM / YY ___	
2. Gender 1. <input type="checkbox"/> Male 2. <input type="checkbox"/> Female	
3. My age is _____ years	
4. What is the closest city or town to where you live? _____	
5. What is the main ethnic group you belong to? (Tick <i>one</i> only) 1. <input type="checkbox"/> New Zealand European / Pakeha 2. <input type="checkbox"/> Maori 3. <input type="checkbox"/> Pacific Peoples 4. <input type="checkbox"/> Chinese 5. <input type="checkbox"/> Indian 6. <input type="checkbox"/> Other Asian 7. <input type="checkbox"/> Other: _____	
About your recovery	
6. Do you have a sponsor now? 1. <input type="checkbox"/> Yes, I speak with or contact my sponsor: <input type="checkbox"/> daily <input type="checkbox"/> weekly <input type="checkbox"/> other _____ 2. <input type="checkbox"/> No I don't have a sponsor.	
7. Do you sponsor others now? 1. <input type="checkbox"/> Yes, I sponsor _____ people. 2. <input type="checkbox"/> No	
8. How often do you go to NA meetings now? (Complete <i>one</i> line only) 1. _____ times a week 2. _____ times a month	
9. Where did you get clean? (tick <i>one</i> only) 1. <input type="checkbox"/> NA 2. <input type="checkbox"/> Treatment. 3. <input type="checkbox"/> Other 4. <input type="checkbox"/> Other 12 Step fellowship please specify _____	
10. Have you started to work the 12 steps? 1. <input type="checkbox"/> Yes 2. <input type="checkbox"/> No	
11. Have you worked on all the 12 steps at least once? 1. <input type="checkbox"/> Yes 2. <input type="checkbox"/> No	
12. What NA service have you been involved with? (Tick <i>all</i> that apply) 1. <input type="checkbox"/> None 2. <input type="checkbox"/> Group service: (secretary, treasurer, coffeemaker, literature person, etc.) 3. <input type="checkbox"/> Area service (inc. sub-committees & GSR) 4. <input type="checkbox"/> Regional service 5. <input type="checkbox"/> World service 7. <input type="checkbox"/> Other: _____	
13. Are you currently doing service for NA? 1. <input type="checkbox"/> Yes 2. <input type="checkbox"/> No	
14. Do you currently attend meetings of any other 12 step fellowships? 1. <input type="checkbox"/> Yes 2. <input type="checkbox"/> No	
15. Do you currently attend other types of addiction counselling or addiction support programs? 1. <input type="checkbox"/> Yes 2. <input type="checkbox"/> No	
16. Have you attended a residential or day treatment programme for addiction? 1. <input type="checkbox"/> No 2. Yes: <input type="checkbox"/> Day <input type="checkbox"/> Residential programme	
17. Please indicate which of the following influenced you the most to come to your first NA meeting. (Tick <i>one</i> only) 1. <input type="checkbox"/> NA member 2. <input type="checkbox"/> NA literature / phoneline / flyer / website 3. <input type="checkbox"/> Treatment facility / health care agency 4. <input type="checkbox"/> NA meeting in prison / corrections facility 5. <input type="checkbox"/> AA member or group 6. <input type="checkbox"/> Family / friend / Nar-Anon member 7. <input type="checkbox"/> Self referred 8. <input type="checkbox"/> Court / probation or parole officer 9. <input type="checkbox"/> Employer or co-worker 10. <input type="checkbox"/> Newspaper, magazine, radio, or TV 11. <input type="checkbox"/> Counsellor, teacher, clergy member 12. <input type="checkbox"/> Other: _____	
Before you came into recovery	
18. How long did you use drugs for in total? 1. _____ years	

PLEASE TURN OVER >>

Fig. A.3: Page 1/2 2008 survey questionnaire

<p>19. What drugs did you use on a regular basis: (Tick <i>all</i> that apply)</p> <ol style="list-style-type: none"> 1. <input type="checkbox"/> Alcohol 2. <input type="checkbox"/> Cannabis (pot, hashish, etc.) 3. <input type="checkbox"/> Methamphetamine (P etc.) 4. <input type="checkbox"/> Cocaine 5. <input type="checkbox"/> Barbiturates (downers, etc.) 6. <input type="checkbox"/> Tranquillisers (Valium, etc.) 7. <input type="checkbox"/> Hallucinogens (LSD, etc.) 8. <input type="checkbox"/> Inhalants (glue, etc.) 9. <input type="checkbox"/> Opiates (heroin, etc.) 10. <input type="checkbox"/> Other stimulants (speed, etc.) 11. <input type="checkbox"/> Methadone <p>Other: _____</p>	<p>26. Main type of paid work you're doing now. (Tick <i>one</i> only)</p> <ol style="list-style-type: none"> 1. <input type="checkbox"/> Unskilled work / labouring 2. <input type="checkbox"/> Clerical work 3. <input type="checkbox"/> Service / sales work / hospitality 4. <input type="checkbox"/> Trades worker 5. <input type="checkbox"/> Agriculture / fishery 6. <input type="checkbox"/> Plant / machine operator 7. <input type="checkbox"/> Craft worker / artist / musician / actor 8. <input type="checkbox"/> Technical 9. <input type="checkbox"/> Manager/administrator/legislator/professional 10. <input type="checkbox"/> Health professional 11. <input type="checkbox"/> Student 12. <input type="checkbox"/> None <p>Other: _____</p>
<p>20. Did you have one drug of choice?</p> <ol style="list-style-type: none"> 1. <input type="checkbox"/> Yes, select number from the list above: _____ 2. <input type="checkbox"/> No 	<p>27. Main type of paid work prior attending NA Please select number from the list above: _____</p>
<p>Other Information</p>	
<p>21. Was there one drug you used the most?</p> <ol style="list-style-type: none"> 1. <input type="checkbox"/> Yes, select number from the list above: _____ 2. <input type="checkbox"/> No 	<p>28. BEFORE your clean date, did you contract / develop a diagnosed health condition?</p> <ol style="list-style-type: none"> 1. Mental Illness: <input type="checkbox"/> None <input type="checkbox"/> Depression <input type="checkbox"/> Bipolar <input type="checkbox"/> Psychosis <input type="checkbox"/> Other 2. Physical Illness: <input type="checkbox"/> None <input type="checkbox"/> Hep. C <input type="checkbox"/> Organ Damage <input type="checkbox"/> HIV <input type="checkbox"/> Other
<p>More about you</p>	
<p>22. Main source of income now. (Tick <i>one</i> only)</p> <ol style="list-style-type: none"> 1. <input type="checkbox"/> Full-time employment (30 hours or more) 2. <input type="checkbox"/> Part-time employment (30 hours or less) 3. <input type="checkbox"/> Self employment, full or part-time 4. <input type="checkbox"/> Retired (Superannuitant) 5. <input type="checkbox"/> Student loan / grant 6. <input type="checkbox"/> Private income 7. <input type="checkbox"/> Unpaid worker (carer, homemaker etc) 8. <input type="checkbox"/> Beneficiary(ACC, sickness, unemployed, DPB) <p>Other: _____</p>	<p>29. Do you still experience the symptoms that resulted in your diagnosis?</p> <ol style="list-style-type: none"> 1. <input type="checkbox"/> No 2. Yes: <input type="checkbox"/> Mental Illness <input type="checkbox"/> Physical Illness
<p>23. Main source of income before attending NA Please select number from the list above: _____</p>	<p>30. SINCE your clean date, did you contract / develop a diagnosed health condition?</p> <ol style="list-style-type: none"> 1. Mental Illness: <input type="checkbox"/> None <input type="checkbox"/> Depression <input type="checkbox"/> Bipolar <input type="checkbox"/> Psychosis <input type="checkbox"/> Other 2. Physical Illness: <input type="checkbox"/> None <input type="checkbox"/> Hep. C <input type="checkbox"/> Organ Damage <input type="checkbox"/> HIV <input type="checkbox"/> Other
<p>24. What is your highest educational qualification now?</p> <ol style="list-style-type: none"> 1. <input type="checkbox"/> No qualification 2. <input type="checkbox"/> School qualification 3. <input type="checkbox"/> Trade / vocational qualification 4. <input type="checkbox"/> Tertiary qualification 5. <input type="checkbox"/> Graduate or postgraduate qualification 	<p>31. Do you have criminal convictions as a result of your drug use?</p> <ol style="list-style-type: none"> 1. <input type="checkbox"/> Yes 2. <input type="checkbox"/> No
<p>25. What was your highest education level prior to attending NA? Please select from the list above: _____</p>	<p>Thank You for your Service NOT COLLECTED? If for any reason this form has not been collected at the meeting please post it to Survey Work Group, PO Box 9051, Wellington by the 29th of November at the latest.</p>

Fig. A.4: Page 2/2 2008 survey questionnaire

How to Run the Survey

1 **Announcement at the beginning of meeting**

- This is "Making Your Recovery Count"- survey week in NA.
- All NA members and anyone else who thinks they may have a drug problem are invited to take part.
- Fill in the anonymous form after the meeting to join this activity.

2 **During the meeting**

- Count heads to see how many people are at the meeting – and record it on the cover sheet

3 **Announcement to start the survey**

- This group has agreed to do the survey.
- All information will be kept anonymous.
- The data analysis will make sure that no-one can be identified from their answers.
- If anyone has trouble, its ok to ask someone else to help you complete the form.
- The answers given will help NA to
 - Understand its membership better
 - "Carry the message" by informing professionals and other agencies about the benefits of NA.
- Please complete **ONLY ONE** survey form this week.
- **Please stay to complete a survey form. It only takes a few minutes.**

4 **Run the survey**

1. Pass out pens.
2. Hand out forms to everyone. Keep count of how many forms you hand out.
3. Collect completed forms, and count how many you get back. (Forms cannot be taken away and completed elsewhere.)
4. Put all forms in the Reply Envelope in the presence of a trusted servant.
5. Seal the envelope and sign across the edge of the seal. Have a trusted servant sign here too.
6. Post the forms as soon as possible. It is ideal if a trusted servant from the meeting goes with you to the post box.

Return forms to: Survey Work Group, PO Box 9051, WELLINGTON.

Thank you for your service in Making Your Recovery Count.

Any queries contact: [REDACTED]

How to Run -App III 5.1.doc

Fig. A.5: How to run the survey

COVER PAGE

Collectors: complete the following at the meeting and just before the survey and return with forms

Total number of addicts at this meeting

Number of who have already attended another meeting ***and completed a form***

Number of forms handed out

Number of **completed** forms returned

Number who refused to fill out a form

Day	
Date	
Time of Meeting	
Name of Group/ Meeting	
Place/Venue	
Name of Town/City	
Name of Area	

It is important that this sheet is completed and returned with all the forms (*including unused forms*) to:
Survey Collections, _____, _____
immediately

Thank you for your service in Making Your Recovery Count.

Any queries contact:

Fig. A.6: Cover page for scribe to note attendance, forms out, forms in, previously completed

Group Announcement***Making Your Recovery Count***

NA is surveying its members in New Zealand.

This group has agreed to be part of the survey.

We will do the survey after our meeting in the week 12-18 November.

With this survey we will:

- **Understand** our membership better, and
- **Carry** the message, because the survey report will help to inform professionals and other agencies about the benefits of NA.

Please note: The survey is anonymous and NOBODY will be identifiable. The survey is voluntary and the questions are easy. It only takes a few minutes.

*Note that this announcement will only be provided to groups that have AGREED at a conscience meeting to join in the survey.

Thank-you for your service in Making Your Recovery Count

Group Announcement 1.doc/pdf

Fig. A.7: Announcement for participating meetings

This form lists all the meetings in your area that requires your responsibility. Some collectors have been identified and their contact details are included.

100

Please forward this information by one of these options:

Fax:

Snail mail: Survey Collections, PO Box 9051, Wellington

BY 10 OCTOBER 2004

Fig. A.8: Collector collection form

How to Run the Survey

1 **Announcement at the beginning of meeting**

- This is "Making Your Recovery Count"- survey week in NA.
- All NA members and anyone else who thinks they may have a drug problem are invited to take part.
- Fill in the anonymous form after the meeting to join this activity.

2 **During the meeting**

- Count heads to see how many people are at the meeting – and record it on the cover sheet

3 **Announcement to start the survey**

- This group has agreed to do the survey.
- All information will be kept anonymous.
- The data analysis will make sure that no-one can be identified from their answers.
- If anyone has trouble, its ok to ask someone else to help you complete the form.
- The answers given will help NA to
 - Understand its membership better
 - "Carry the message" by informing professionals and other agencies about the benefits of NA.
- Please complete ONLY ONE survey form this week.
- **Please stay to complete a survey form. It only takes a few minutes.**

4 **Run the survey**

1. Pass out pens.
2. Hand out forms to everyone. Keep count of how many forms you hand out.
3. Collect completed forms, and count how many you get back. (Forms cannot be taken away and completed elsewhere.)
4. Put all forms in the Reply Envelope in the presence of a trusted servant.
5. Seal the envelope and sign across the edge of the seal. Have a trusted servant sign here too.
6. Post the forms as soon as possible. It is ideal if a trusted servant from the meeting goes with you to the post box.

Return forms to: Survey Work Group, PO Box 9051, WELLINGTON.

Thank you for your service in Making Your Recovery Count.

Any queries contact: [REDACTED]

How to Run -App III 5.1.doc

Fig. A.9: Information on how to run the survey

B. ADDITIONAL GRAPHS

This appendix contains additional graphs of the comparison of the matched and unmatched groups for 2004 and 2008.

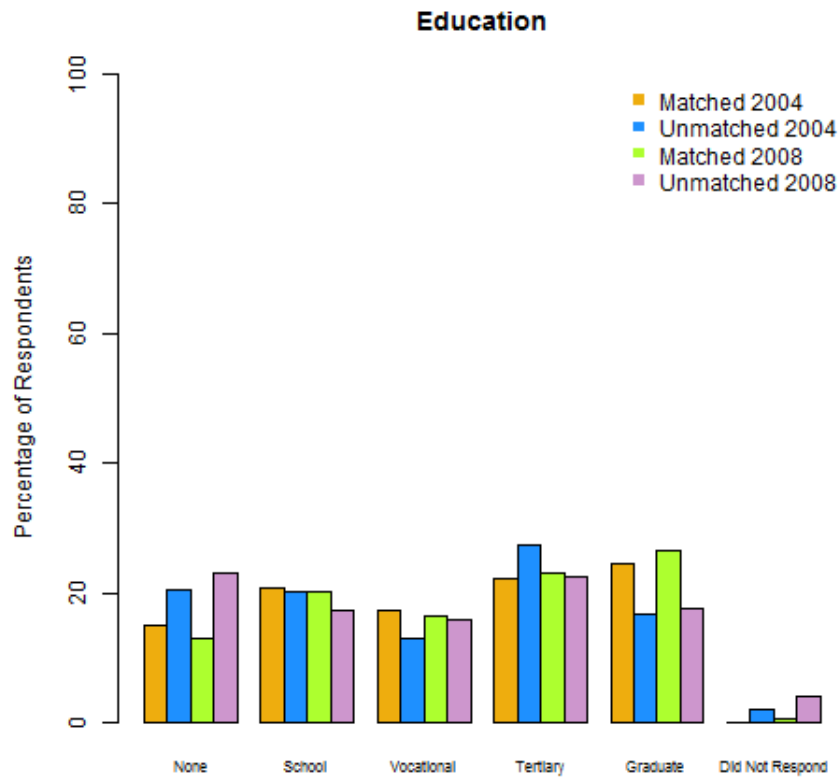


Fig. B.1: Distribution of individuals education level

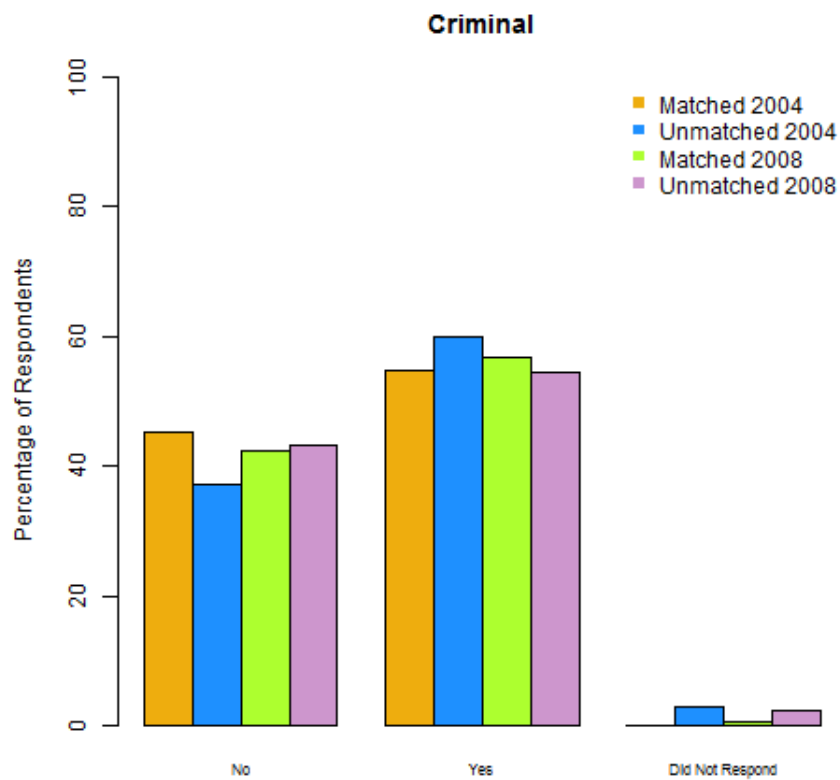


Fig. B.2: Distribution of if individual holds criminal record

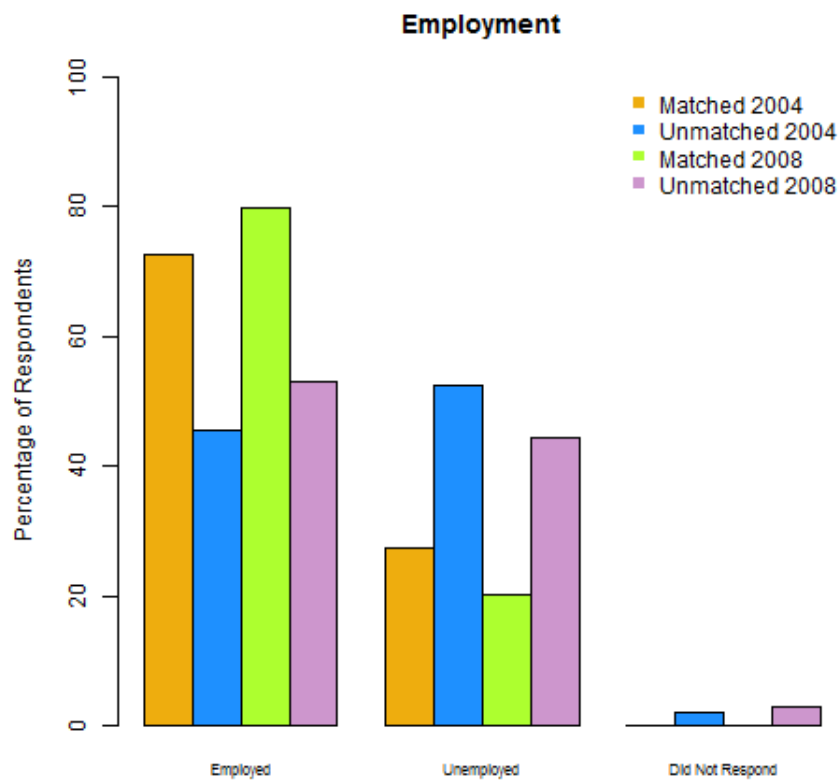


Fig. B.3: Distribution of individuals employment status

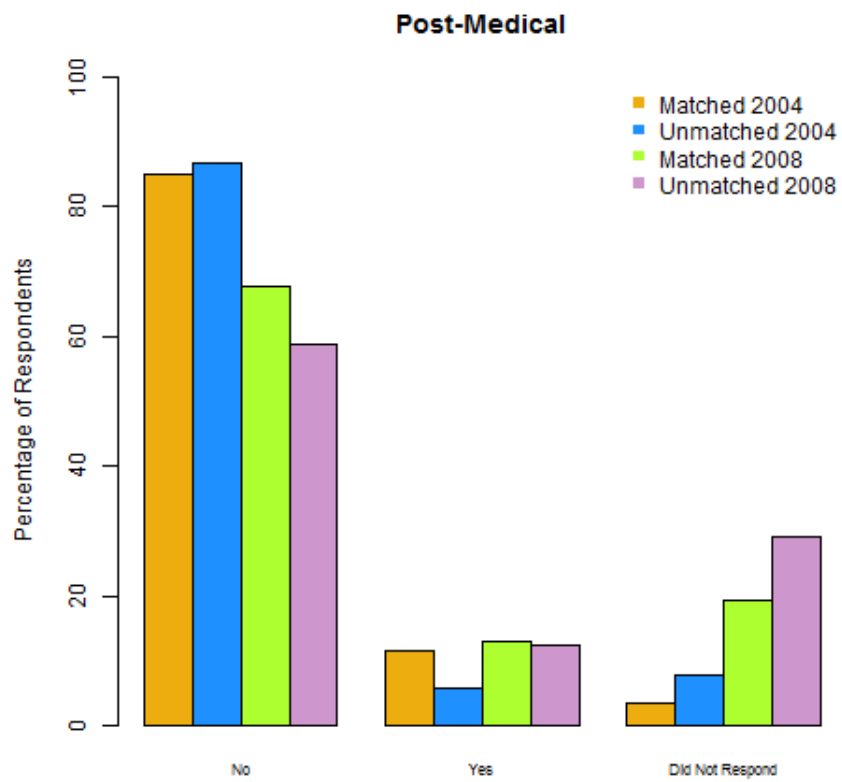


Fig. B.4: Distribution of individuals medical state after joining NA

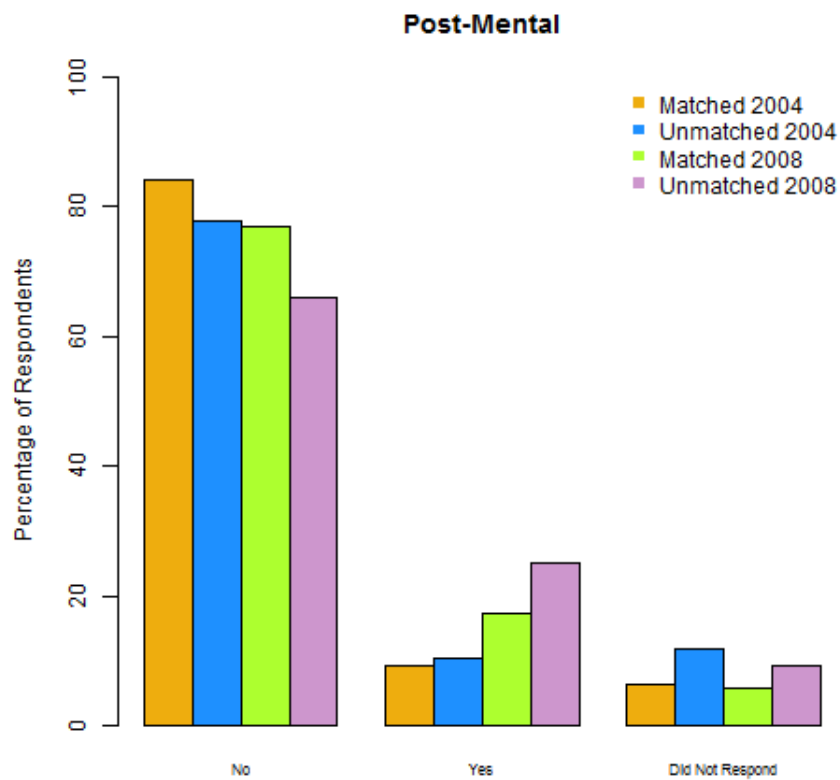


Fig. B.5: Distribution of individuals mental state after joining NA